EFFICIENT AND DISTRIBUTED COMPUTATIONAL METHODS FOR COMPLEX SYSTEMS

A Dissertation
Presented to
The Academic Faculty

By

Yuchen Zheng

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and System Engineering

Georgia Institute of Technology

May 2018

EFFICIENT AND DISTRIBUTED COMPUTATIONAL METHODS FOR COMPLEX SYSTEMS

Approved by:

Dr. Nicoleta Serban
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Yao Xie
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Huan Xu
H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology

Dr. Anne Fitzpatrick
School of Medicine
Emory Univeristy

Dr. Ilbin Lee
Alberta School of Business
University of Alberta

Date Approved: March 30, 2018

Remember to look up at the stars and not down at your feet. Try to make sense of what you see and wonder about what makes the universe exist. Be curious.

Stephen Hawking

I dedicate this humble work to my amazing parents and my lovely wife, who have been a great source of love, inspiration and encouragement. I also dedicate this thesis to my advisor Dr. Nicoleta Serban, who nutured a childlike curiosity in many things.

ACKNOWLEDGEMENTS

There are many people I'd like to thank both for completion of this Thesis and throughout my Ph.D. studies. First, I would like to thank my advisor, Dr. Nicoleta Serban, for all her support, guidance and caring for the past five years, both academically and personally. Dr. Serban took me in as a undergraduate researcher since my Junior year, and we've collaborated ever since. Her broad knowledge, unique way of stimulating my interest and high standard have helped me tremendously. Without her, none of the research in this thesis would be possible. I would also like to thank my committee members, Dr. Yao Xie, Dr. Huan Xu, Dr. Ilbin Lee, and Dr. Anne Fitzpatrick for taking time from their busy schedule to be in my committee and help me improve this thesis.

I would like to thank my collaborators, Dr. Ross Hilton, Dr. Anne Fitzpatrick, and Dr. Ilbin Lee. Ross and I collaborated on a few projects, and he played an important role in my early Ph.D. research. Dr. Fitzpatrick has provided significant domain knowledge for our papers related to pediatric Asthma. Although I have only worked with Dr. Lee for one year, it was the most productive year and we have accomplished a lot together. Dr. Lee's high standard, attention to detail and knowledge on optimization made it possible for us to explore many interesting research topics. I would also like to thank Richard Starr, Matt Sanders and Paul Diedrich at IPaT who provided the data safeguard and technology infrastructure for the Medicaid data. I would like to thank two of the undergraduate researchers I worked with, Preston and Henry, who provided great help at moment's notice.

In addition, I would like to thank my friends, Yuan, Zihao, Sinan, Henry, to name a few, who have been very supportive and are great sources of encouragement. Zihao and I are having a never ending friendship since day one of undergraduate. Yuan became an important part of our lives, but later ditched us for Seattle. Thanks bro.

None of this is possible without the support and love from my parents, who have been a beacon of inspiration and motivation for me throughout these years. Their dedication, sac-

rifice and love grant me the very opportunity to finish this thesis and to earn this prestigious degree.

Most importantly, I would like to thank the love of my life - my wife Zhaocheng. Throughout the past seven years of mostly ups and rarely downs, she has always been there for me. Whenever I got frustrated or ran into a dead end in research, I would take a break and spend some time with her. The problems always resolve themselves afterwards, like a magic. Without her, I wouldn't imagine how I could finish any of these.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Many statistical inference problems for large-scale complex systems involve using analytical tools, such as statistics, mathematical optimization and machine learning algorithms. Due to the explosion in size and complexity of modern datasets, traditional ways of modeling on a single computing node are no longer scalable. Some researchers have suggested that even highly complex and structured problems characterized by large datasets may fail to be explored even with relatively simple models.[1] Solving a large-scale problem using serial computing can be computationally challenging due to compute load and memory usage. In addition, data may be stored in a decentralized fashion, and communicating data to a centralized location can be wasteful, and may cause privacy issues. [2][3] One remedy to such challenges is to utilize the power of distributed computing and distributed data storage. A serial algorithm is therefore decomposed into many subroutines and a large problem is split into many sub-problems that can be solved concurrently using different computing nodes. In cases where there is little or no dependency or need for communication between different parallel tasks, a serial algorithm can be converted into distributed algorithm in an *embarrassingly parallel* fashion, such as parallelizing the summation of an array of numbers, or the Maximum Likelihood Estimation of independent samples. However, many real applications and algorithms have inherent structures that prevents such a straightforward decomposition, such as *spatial coupling* for optimization problem and modeling interdependent samples among others. One main approach considered in this thesis is addressing the computational scalability in complex systems using distributed computing.

We demonstrate the importance of innovation in computational statistics with rigorous analysis of large medical datasets. Data in healthcare are generated at every patient's encounter with the healthcare system, at every implementation of medical processes, with

every decision made by healthcare organizations, and with every policy implementation in the healthcare ecosystem, resulting in billions of data points every day. Every patient in any medical setting generates an invaluable data point that can contribute to understanding what works, for whom and where. Developing analytical methods to translate these types of data into meaningful knowledge is crucial to help us better understand behavior patterns in seeking care and adherence to recommended care guidelines, and derive knowledge for decision support. Other types of complex systems include transportation, social network, logistics among others. Although these problems arise in diverse application domains, they share some important characteristics and challenges. First, while different components of the system are mostly heterogeneous, there are much homogeneity in characteristics that can be explored. Second, interactions between different components interdependently give rise to collaborative patterns in the system. Third, the quality of the data is polluted by unquantifiable random noise or errors. Fourth, the datasets are often extremely large in scale and dimensionality, since advanced technology allows us to collect and store every detailed information about each sample. Therefore, it has become of central importance to develop scalable computational algorithms that can describe, profile, and model these systems and help make robust decisions.

In this dissertation thesis, we propose several computational efficient methods that model complex systems in different settings. In Chapter 2, we introduce a framework for analyzing and visualizing the healthcare utilization for millions of children, with a focus on pediatric asthma. Using individual-level claims data across 10 southeast states for the Medicaid system, we model the heterogeneity in patients' multi-year longitudinal utilization patterns via mixture Markov renewal processes. In Chapter 3, we introduce a regularized optimization approach to control the trade-off between optimality and sensitivity of the solution to large-scale optimization problems that has intrinsic spatial structure among decision variables. We illustrate the proposed approach using a specific application in health care access measurement, in which a smooth solution that is robust to perturbations

of model parameter leads to reliable decision-making. In Chapter 4, we propose a novel method to find a partition of decision variables for decomposing large-scale optimization problems, focusing on minimizing the number of dualized constraints. We present an improved variation of the distributed sub-gradient method using block dual decomposition. In Chapter 5, we develop a computationally tractable algorithm for clustering spatially dependent data using the EM algorithm, and cluster the prevalence of chronic conditions among children with Medicaid in the entire United States at the community level. The implementation of the spatial clustering approach relies on distributed computing to overcome the computational effort needed to perform the analysis.

CHAPTER 2

UNCOVERING LONGITUDINAL HEALTHCARE BEHAVIORS FOR MILLIONS OF MEDICAID ENROLLEES: A MULTI-STATE COMPARISON OF PEDIATRIC ASTHMA UTILIZATION AND COST

In this chapter, we introduce a framework for analyzing and visualizing healthcare utilization and cost for millions of children, with a focus on pediatric asthma, one of the major chronic respiratory conditions. We use the 2005-2012 Medicaid Analytic Extract claims for 10 southeast states. The study population consists of Medicaid-enrolled children with persistent asthma. We translate multi-year, individual-level medical claims into sequences of discrete utilization events, which are modeled using Markov renewal processes and model-based clustering. Network analysis is used to visualize utilization profiles. The method is general, allowing the study of other chronic conditions. The study population consists of 1.5 million children with persistent asthma. All states have profiles with high probability of asthma controller medication, as large as 60.6% and 90.2% of the state study population. The probability of consecutive asthma controller prescriptions ranges between 0.75 and 0.95. All states have utilization profiles with uncontrolled asthma with between 4.5% and 22.9% of the state study population. The probability for controller medication is larger than for short-term medication after a physician visit but not after an emergency department (ED) visit or hospitalization. Transition from ED or hospitalization generally has a lower probability into physician office (between 0.11 and 0.38) than into ED or hospitalization (between 0.20 and 0.59). The highest level of adherence is with respect to asthma controller medication. In most profiles, children who take medication do so regularly. The lowest level of compliance is with follow-up physician office visits after an (ED) encounter or hospitalization. Finally, all states have a proportion of children who have uncontrolled asthma, meaning they do not take controller medication while they do have severe out-

comes.

2.1 Introduction

Data in healthcare are generated at every patient's encounter with the healthcare system, at every implementation of medical processes, with every decision made by healthcare organizations, and with every policy implementation in the healthcare ecosystem, resulting in billions of data points every day. Every patient in any medical setting generates an invaluable data point that can contribute to understanding what works, for who and where.

One health-related information technology (IT) that has provided substantive opportunities to study healthcare across large populations and many years is the medical claims system. Information coded in claims data is standardized to a great extent [4], hence making such data amenable to large scale studies. Developing methods to translate medical claims data into meaningful information is the first crucial step in medical decision making. Further development of adaptive and scalable data mining and statistical methods provide the means for analyzing these data. However, there are a series of challenges associated with mining data derived from medical claims, including the derivation of knowledge for decision support while maintaining computational efficiency and complying with privacy safeguards.

In this study, we propose a method that translates medical claims data into individual-level utilization sequences, longitudinally over mutliple years, and that models heterogenity in individual-level healthcare utilization using mixture Markov renewal processes. The method combines the benefits of network analysis and model-based clustering for discrete event sequences to also provide visual summaries of underlying utilization profiles. Thus one contribution is a model-based data mining algorithm that has the ability to scale to massive data while producing meaningful stochastic networks that can then be used in decision

support through visualization and simulation. The second contribution is the application of the modeling approach to derive inferences on utilization behaviors from highly-sensitive, large patient-level claims data.

We demonstrate the applicability of the proposed methods in health policy and medical decision making in drawing inferences on longitudinal utilization for pediatic asthma healtchare. Asthma is the most common respiratory chronic disease in children [5]. More than 10 million children have had asthma in their lifetime [6], with 42.9% classified as uncontrolled [7]. While asthma cannot be cured, with the appropriate medication and treatment plan, its symptoms can be controlled [8][9]. Controlling asthma is important for children since it can prevent damage to growing lungs but can also improve their quality of life and potentially reduce the cost of care by preempting severe health outcomes [10][11][12][13]. The U.S. Centers for Disease Control and Prevention has identified pediatric asthma as a priority condition for intervention [14].

This study is the first to uncover multi-year longitudinal utilization for pediatric asthma care using individual-level claims data across 10 states (Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas) for the Medicaid system. We focus on the southeast due to the great disparities and poor health outcomes there[15]. Some southeastern states have among the highest expenditures, such as Georgia [16], of all states [17].

2.2 Methods

2.2.1 Data Sources

The main data source is the Medicaid Analytical Extract (MAX) medical claims data acquired from the Centers of Medicare and Medicaid Services (CMS), consisting of identifiable individual-level claims data for all Medicaid-enrolled beneficiaries. The MAX dataset

consists of billions of claims records across more than 10 million children enrolled annually in the 10 selected states.We provide the main data elements we extracted from the MAX files in Web-Appendix A.

2.2.2 Study Population

The study population consists of Medicaid-enrolled children ages 4-18 with an asthma-related diagnosis. (We exclude the age group 0-3 from this study because of the difficulty and inaccuracy of diagnosing asthma at this age.) Consistent with standard modifications [17] to the Healthcare Effectiveness Data and Information Set (HEDIS) measures defined by the National Committee for Quality Assurance (NCQA) [18], we define patients with asthma as those who meet one of the criteria:

- At least two visits to physician's office,
- At least two asthma controlled medication prescriptions,
- One emergency room visit or hospitalization with a diagnosis of asthma in addition to at least another visit and/or medication prescription.

These modified filtering criteria help to avoid including patients with incorrect diagnoses of asthma.To capture longitudinal utilization behaviors, we only consider those patients that qualify for Medicaid for at least four of the eight years between 2005 and 2012.

2.2.3 Translating Claims into Indivudal-level Utilization Sequences

We filter the claims based on the ICD-9 diagnosis codes and date of birth to obtain the study population. The MAX claims are structured into inpatient care (IP), long-term care (LT), other care including outpatient services (OT), patient summary (PS) and prescription claim summary (RX) files. Included for each claim are data entries specifying the date of service, the Medicaid Statistical Information System identification (MSIS ID) of each Medicaid

enrolee, the International Classification of Diseases, Ninth Revision (ICD-9) codes for diagnosis or services provided, and the type and place of services rendered. We use the IP and OT files to determine the visits to a specific provider type, and the RX file to determine the medication type and date of the prescription being filled. We abbreviate the derived event types as follows: emergency room visits (ER), hospitalizations (HO), physician's office visits and clinic visits (PO), ashtma short-term medication (ASM) and asthma controlled medication (ACM). The first three event types are derived from the place of service and type of service codes of the claims code in the IP and OT files. Starting with a dataset including a total of more than 40 millions of claims across the 10 states, we derive utilization data consisting of 24 million total events.

The output of this translation step are individual-level sequences of utilization events with corresponding event time stamps. For example, consider a patient who visits the emergency room for an asthma attack on January 1st, 2005, who subsequently receives a prescription for an inhaler which she fills one month later, along with a referral to a primary care physician. She then visits the same physician and refills her asthma prescriptions occur at 3 month intervals. The sequence then is given by (ER,ACM,PO,PO,ACM) with time stamps scaled to one-year intervals: (0.08, 0.25,0.50,0.75).

2.2.4 Model-based Clustering of Utilization Sequences

Modeling Utilization Sequences

We model patient-level utilization in the form of sequential event data over a period of time, where we consider both the order of the events (e.g. an ER visit precedes a PO visit) and the timing between events (e.g. the expected time between two PO visits).

A simple, but useful, model for summarizing time-ordered events with varying time intervals between events is the Markov renewal process (MRP) (Foufoula-Georgiou and Lettenmaier 1987). The MRP is the continuous-time analog of a discrete-time Markov chain (DTMC).

Let $\vec{X} = (X_1, x_2, ..., x_L)$ the sequence of events and $\vec{T} = (T_1, T_2, ..., T_L)$ the set of "arrival" times, where L is the length of the patient healthcare utilization sequence. Let s_i, $i \in \{1, ..., S\}$ be all possible states in the sequences of events, in our case, they are ASM, ACM, ER, HO and PO, where S is the number of states, in our case, S=5.

In an MRP, the sequence $\vec{X}$ is itself a DTMC, with corresponding transition matrix P where P_{ij} is the transition probability between states s_i and s_j and $\sum_{j=1}^{S} P_{ij} = 1$. For example, the transition probability from ER to PO is the probability that a patient receives care in a physician's office after an emergency room encounter. We estimate the transition probabilities P_{ij} using maximum likelihood estimation as presented in Appendix B.

Now we define the distribution for the sequence of interarrival times $\tau_l = T_{l+1} - T_l$. We assume that for each pair $i, j \in \{1, ...S\}$, the distribution of the interarrival times between states s_i and s_j is an exponential distribution with rate parameter λ_{ij}. To estimate the rate parameters we use maximum likelihood estimation provided in Appendix B.

The output of the MRP consists of estimated transition probability and inter-event time matrices, specifying the transition probability and inter-event time expectation for each event type or state pair. Both output matrices are 5-by-5 matrices, where a cell in the transition probability matrix and in the inter-event time matrix respectively corresponds to the probability and the expected time of a patient with asthma to transition from one event type to the same event type or to a different event type. The matrix is not symmetric, since the

transition probability from ER to PO may not be equal to the probability from PO to ER, for example. In order to capture the probability distribution of the entire patient sequence we account for the choice of the first and last events by including left and right censor events. We provide an illustration of an MRP output in Appendix B.

Clustering Analysis of Utilization Sequences

We complement the MRP modeling of the utilization sequences with a clustering of the utilization sequences. We assume that the cluster membership is a latent variable with a multinomial distribution, hence the resulting model is a mixture MRP. The MRP clustering algorithm simultaneously estimates the MRP model parameters and clusters patients into distinct utilization profiles.

We employ the estimation maximization (EM) with a hierarchical tree-based algorithm to derive the clustering output. In the first step, the algorithm searches for a division that maximizes the Bayesian information criterion score using the Kullback Leibler distance between the individual patient distributions and population distributions given a set of candidate divisions. We then use this candidate division as the initialization for the EM algorithm where patients move to clusters such that to maximize each individual posterior likelihood. After assigning patients to a new cluster, the parameters are re-estimated given current cluster membership via maximum likelihood estimation of both the transition probabilities and inter-event time distributions.

The computational complexity of our algorithm is O(n log n). The primary computational steps involved in fitting a patient sequence to an MRP rely on simple counting and averaging, while the computation of posterior likelihood relies on multiplication. The sorting step of the posterior likelihoods in determining the clustering membership is the most

computationally expensive with order O(n log n). All other computations are of order O(n).

Once we identify a clustering tree using the approach described below, we further re-cluster nodes that have similar utilization patterns in terms of both the event types or states with higher probabilities as well as in terms of the expected interarrival times. This allows a sparser representation of the utilization profiles. We provide the details of the model structure and the model estimation in Appendix B.

Utilization Profile Visualization

By employing stochastic models for clustering utilization sequences we can further derive stochastic provider networks via the transition matrices, allowing for visualization of the utilization behaviors as networks across different healthcare types or states. The primary inputs for the stochastic provider networks are the transition matrices. Specifically, the five event types, ASM, ACM, ER, HO and PO, are the nodes in a directed graph. The directed edges represent transition probabilities between two event types, for example, the transition from the ER to a PO visit. For a simplified representation, the networks only include nodes such that a total of 90% of volume is represented. We use different types of arcs for different levels of transition probabilities to better identify nodes that are most visited within each profile.

We provide a graphical representation of the translation of one MRP as a network in Figure 2.1. For each state and cluster our algorithm calculates the raw transition matrices: the probability transition matrix which contains the probability of a follow-up visit from one provider type to another, and the average interarrival time matrix containing the average length of time between visits in months. One can ascertain from the raw transition probabilities that the only events with a high probability of visits are the physician's of-

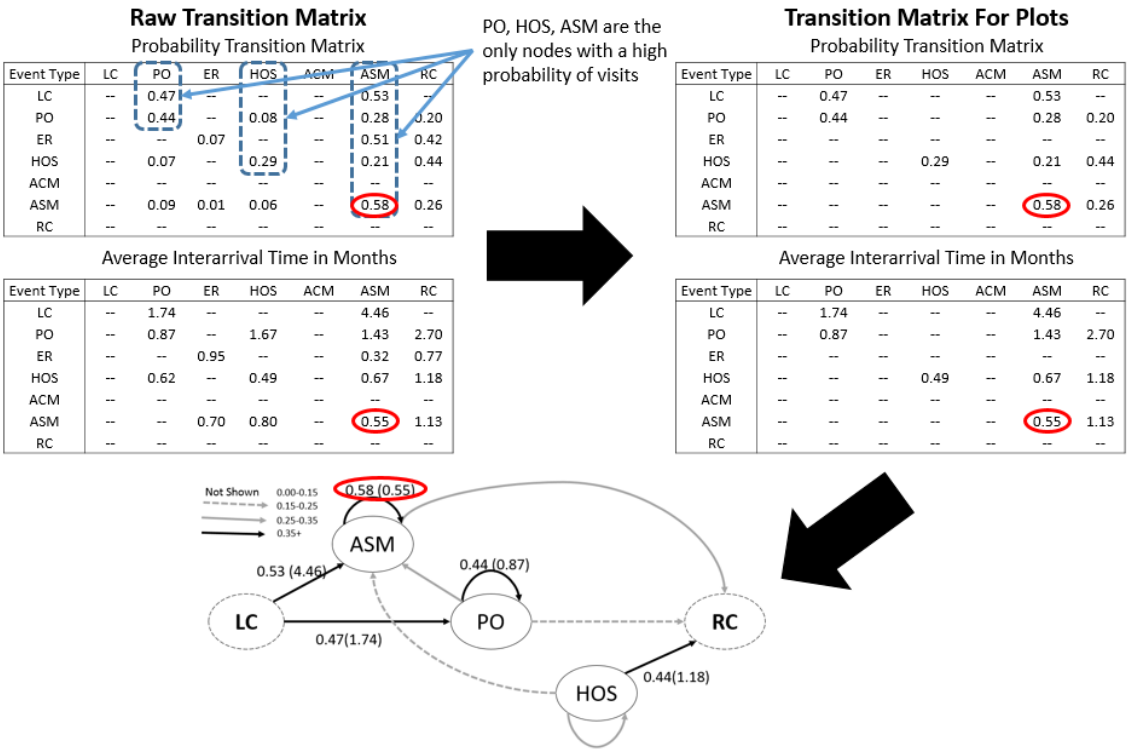


Figure 2.1: Illustrative example of translating MRPs into utilization network graphs.

fice, a short term prescription fill and hospitalization. Therefore, our the transition matrices used as inputs for the plots contain only these three provider types. The nodes ASM, PO and HOS are circled in our provider networks. Consider the transition from ASM back to ASM. The probability of a repeated ASM visit after a ASM visit is 0.58 with an average time between visits of 0.55 months. The directed edges represent the transition probability between events and the average interarrival time, measured in months, between two consecutive events in parentheses. The left censor (LC) and right censor (RC) nodes represent the beginning and end of the study time period, January 1st, 2005, and December 31st, 2012, respectively. These two nodes are surrounded by dashed circles to differentiate them from actual healthcare events. The three styles and shading schemes of the lines corresponding to transitions between providers help the reader to visualize the high-probability patterns through the network.

Table 2.1: Overall utilization summary for each of the 10 states, sorted by the total number of patients considered for each state. The average number of events, as well as averages of each event types(ACM, ASM, ER, HO, PO) are per member-year.

State	# Patients	avgEvents	avgACM	avgASM	avgER	avgHO	avgPO
MS	40,147	1.94	1.90	0.26	0.08	0.09	0.25
SC	65,175	1.84	1.95	0.37	0.06	0.08	0.32
AR	71,369	2.27	2.07	0.51	0.00	0.10	0.31
AL	104,531	2.03	2.08	0.47	0.04	0.09	0.37
FL	122,667	1.86	1.54	0.32	0.06	0.05	0.53
TN	137,148	1.74	2.04	0.38	0.05	0.06	0.31
GA	137,519	1.99	1.54	0.50	0.06	0.09	0.38
LA	142,608	2.29	1.63	0.34	0.07	0.05	0.26
NC	157,011	1.79	2.17	0.48	0.06	0.08	0.43
TX	476,345	1.76	1.66	0.40	0.04	0.03	0.31
AVG		2.74	1.86	0.40	0.05	0.07	0.35
STD		0.28	0.24	0.08	0.02	0.02	0.08

2.3 Results

2.3.1 Data Summaries

The target population of this study consists of 1.5 million patients with persistent asthma who contributed to 24 million events in the 10 states. Detailed statistics of event types per state are in Table 2.1.

The number of patients per state ranges from 40,000 in South Carolina, to 476,000 in Texas. Since different patients will have different numbers of Medicaid eligibility months per year, we normalize the event counts by member-year: We divide the counts by the number of eligibility months for each patient and multiply by 12. On average, patients included in the analysis are enrolled in Medicaid for 74 months (6.1 years) from 2005 to 2012. The average number of events per member-year is 2.74; in high utilization states such as NC, patients have 3.22 events per member-year, and in low utilization states such as LA, patients have 2.36 events per member-year.

Out of the 24 million total asthma events, 13% are PO visits and 83% are RX events, including ACM and ASM. The numbers of PO and RX events per member-year across the study population are 0.35(0.08) and 2.26(0.28), respectively. NC have significantly higher numbers of RX events at 2.65 per member-year compared with FL, with only 1.86 RX events per member-year. FL patients visit the PO most frequently, with 0.53 per member-year, whereas MS visit the PO least frequently, with 0.25 per member-year.

ER and HO events are less frequent than PO and RX among all states. The average number of ER and HO events per member-year across the entire study population are 0.05(0.02) and 0.07(0.02), respectively. AR has primarily HO events with an extremely small number of ER events. The lowest aggregated rate of ER and HO events occurs in TX (0.07) and the highest occurs in MS (0.17).

2.3.2 Clustering of Utilization Sequences & Visualization of the Utilization Profiles

We first describe the underlying utilization patterns for medication arising across all utilization profiles for the 10 states:

- **ACM** – high probability controller medication but low probability short-term medication;
- **ASM** – high probability short-term medication but low probability controller medication;
- **ACM** $\leftarrow$ **ASM** – high probability controller and short-term medication with high probability link between the two medications;
- **ACM/ASM** – high probability controller and short-term medication with low probability link between the two medications.

Table 2.2: Characterization of utilization profiles for each state along with the precentage of patients for each cluster in pharanthesis.

State	Profile 1 Pop %	Profile 2 Pop %	Profile 3 Pop %	Profile 4 Pop %	Profile 5 Pop %
AL	ACM<–ASM + (PO) 80.70%	ACM<–ASM + (PO, HOS) 4.30%	ASM +(PO, HOS) 8.20%	ASM +(PO, HOS,ER) 6.80%	
AR	ACM<–ASM + (PO) 35.70%	ACM 24.90%	ACM/ASM +(PO) 20.60%	ASM +(HOS) 5.20%	ACM/ASM +(PO,HOS) 13%
FL	ACM/ASM +(PO) 32%	ACM<–ASM +(PO) 61.60%	ACM +(PO ,ER,HOS) 6.40%		
GA	ACM<–ASM +(PO, HOS) 72.20%	ASM +(PO) 13.50%	 + (PO) 2.40%	ASM +(PO, ER,HOS) 11.90%	
LA	ACM 5.50%	ACM<–ASM + (PO) 14.80%	ACM/ASM + (PO) 6.10%	ACM<–ASM + (PO, ER) 60.30%	ASM+(PO, ER,HOS) 13.40%
MS	ACM +(PO) 8.20%	ACM/ASM +(PO) 5.50%	ACM<–ASM +(PO,HOS) 63%	ASM +(PO, ER,HOS) 11.20%	ASM +(PO,ER) 0.60%
NC	ACM<–ASM + (PO) 89.70%	+(PO, ER , HOS) 2.60%	ASM+(PO, ER, HOS) 7.60%		
SC	ACM<–ASM + (PO) 70.10%	ASM +(PO) 8.20%	ASM+(PO, ER,HOS) 4%	ASM +(PO, ER,HOS) 14%	
TN	ACM 8%	ACM<–ASM + (PO) 64%	ASM + (PO) 3.10%	ASM + (PO, ER, HOS) 6.50%	ACM<–ASM + (PO, HOS) 18.40%
TX	ACM 8.60%	ACM +(PO) 10%	ACM<–ASM + (PO) 71.60%	ACM/ASM +(PO) 5.10%	ASM + (PO, ER, HOS) 4.70%

Each utilization profile is characterized by one of the four medication patterns described above or no medication, in addition to other types of care, including PO, ER and HOS, depending on which of the event types are present in the networks. For example, the profile ASM+PO describes utilization of short term medication and physician office for asthma care. The profile ASM+(PO,ER,HOS) described utilization of short term medication along with physician office, emergency department and hospitalizations. Table 2.2 provides the description of the profiles for all 10 states. Figure 2.2 is the utilization network for Alabama as an illustrative examples. Figures in Appendix C are the utilization networks for the rest of the states.

Common features across all profiles in the 10 states are:

- All states have profiles with high probability ACM (ACM or ACM $\leftarrow$ ASM). The percentages of the study population across the nine states are: 85%-AL, 60.6%-AR, 61.6%-FL, 72.2%-GA, 80.6%-LA, 71.2%-MS, 89.7%-NC, 65.2%-SC, 72%-TN, and 90.2%-TX.

- All states have one or more profiles with severe outcomes (ER or/and HOS).

- For those profiles with ACM $\leftarrow$ ASM and severe outcomes, all except one profile in MS have a high probability link to ACM; the percentages of the study population among such profiles are: 4.3%-AL, 72.2%-GA, 60.3%-LA, 18.4%-TN.

- For those profiles with different medication patterns (ACM, ACM/ASM or ASM) and severe outcomes, the link from ER or HOS to ACM has low probbility; the percentages of the study population among such profiles are: 15%-AL, 16.2%-AR, 6.4%-FL, 11.9%-GA, 13.4%-LA, 11.8%-MS, 10.2%-NC, 22.9%-SC, 6.5%-TN and 4.7%-TX.

Table 2.3 provides ranges of the transition probabilities across all profiles by state for links pertinent to recommended guidelines for asthma care.

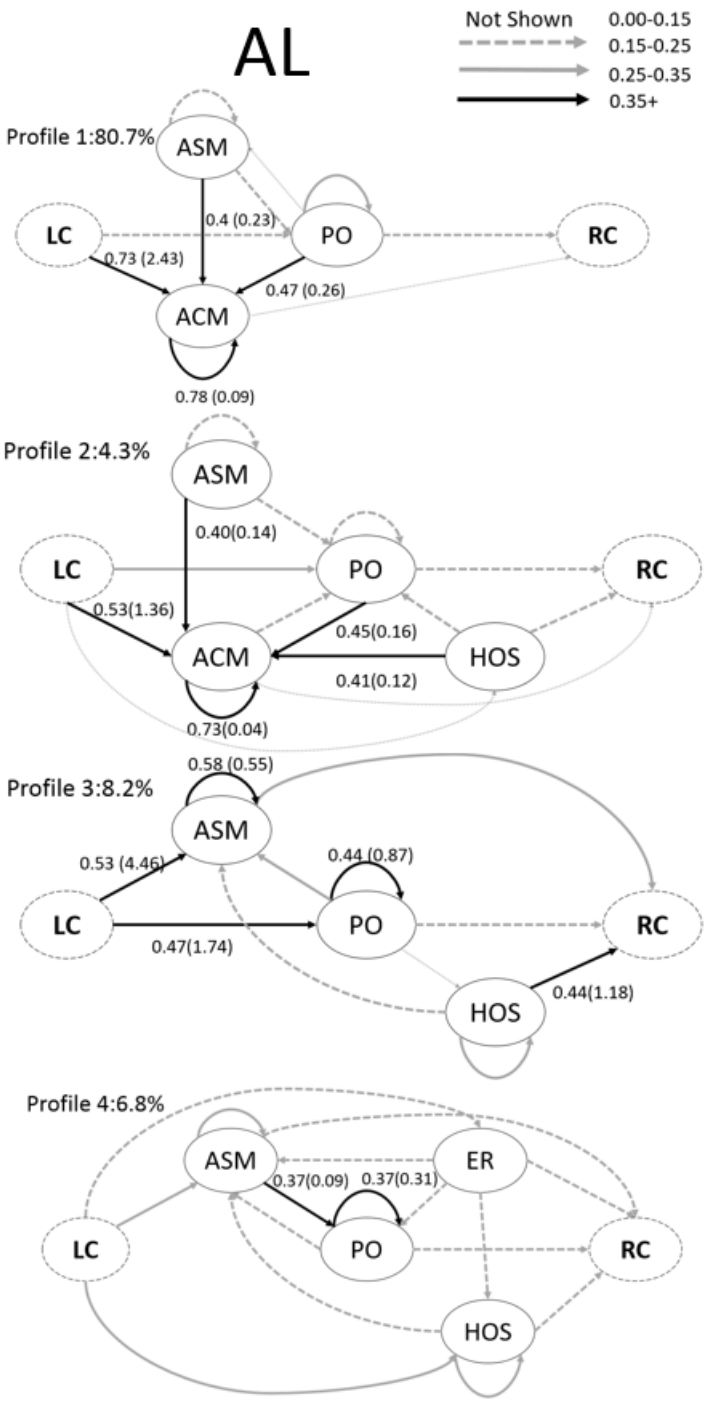


Figure 2.2: Network graphs of etimated utilization profiles of AL. Transition probabilites are given on each edge along with the average interarrival times measured in months in parentheses. Some important edges with probability less than 0.15 are displayed in gray dotted lines.

Table 2.3: The range (minimum and maximum) of probabilities for different links between events across the utilization profiles of all the states.

State	ER/HO ->PO	ER/HOS ->ER/HOS	PO->ACM	PO->ASM	ER/HOS ->ACM	ER/HOS ->ASM
AL	(0.00, 0.19)	(0.07, 0.34)	(0.45, 0.47)	(0.00, 0.28)	(0.32, 0.32)	(0.05, 0.51)
AR	(0.00, 0.11)	(0.00, 0.58)	(0.21, 0.46)	(0.14, 0.26)	(0.00, 0.00)	(0.00, 0.33)
FL	(0.23, 0.24)	(0.06, 0.26)	(0.23, 0.44)	(0.13, 0.18)	(0.14, 0.19)	(0.00, 0.00)
GA	(0.14, 0.18)	(0.00, 0.34)	(0.11, 0.40)	(0.11, 0.20)	(0.28, 0.28)	(0.12, 0.21)
LA	(0.12, 0.13)	(0.08, 0.25)	(0.05, 0.47)	(0.10, 0.24)	(0.05, 0.41)	(0.13, 0.21)
MS	(0.03, 0.13)	(0.08, 0.59)	(0.08, 0.51)	(0.09, 0.23)	(0.07, 0.40)	(0.05, 0.21)
NC	(0.11, 0.24)	(0.09, 0.36)	(0.41, 0.41)	(0.14, 0.15)	(0.00, 0.00)	(0.13, 0.17)
SC	(0.00, 0.19)	(0.00, 0.50)	(0.03, 0.52)	(0.05, 0.20)	(0.09, 0.09)	(0.18, 0.23)
TN	(0.16, 0.38)	(0.00, 0.21)	(0.42, 0.42)	(0.01, 0.42)	(0.22, 0.22)	(0.00, 0.20)
TX	(0.11, 0.12)	(0.00, 0.20)	(0.02, 0.47)	(0.14, 0.29)	(0.00, 0.00)	(0.24, 0.31)

- Follow-up visit after a severe outcome: The probabilities for the links ER/HOS $\rightarrow$ PO are generally lower than for the links ER/HOS $\rightarrow$ ER/HOS; the maximum value across the profiles with each state for ER/HOS $\rightarrow$ PO varies between 0.11 (AR) and 0.38 (TN) and for ER/HO $\rightarrow$ ER/HOS varies between 0.20 (TX) and 0.59 (MS).

- Prescription (re)fill of controller versus short term medication after ER, HOS or PO event: The probability for controller medication after a PO visit are larger than for short-term medication. The probability for controller medication after a ER or HOS is lower than for short-term medication can be lower for some states, including AL, AR, MS, NC, SC and TX.

- Consecutive prescriptions of controller medication have higher probabilities with shorter frequency between (re)fills than those for short-term medication across all states. The probability of an asthma controller medication consecutive prescriptions ranges between 0.75 and 0.95. The frequency of the refill ranges from 0.04 (approximately 2 weeks) to 0.17 (approximately 2 months), with one exception where it is 0.53 or approximately 6 months.

2.4 Discussion

This study brings to bear substantive contributions to the knowledge and decision making on the healthcare delivery for pediatric asthma. It is the first study to draw inferences on patient-level healthcare pathways across approximately 1.5 million children with persistent asthma, the largest study to date. Second, it provides a rigorous approach to model multi-year longitudinal utilization, accounting for both the order of the events and the inter-event time using stochastic modeling, thus not only estimating the frequency of care events, as is common in the existing healthcare utilization research, but also the transition from one care event to another and the expected time between events, relevant in making inferences on the compliance with recommended care. Third, this study identifies similarities and dissimilarities in healthcare utilization across 10 states.

According to the National Heart Blood and Lung Institute [19], the guidelines for asthma treatment specify the use of controller medication for children with persistent asthma. For all 10 states, controller medication is the most prevalent care event, with North Carolina being the highest utilizer and Florida and Georgia the lowest utilizers of controller medication. Importantly, those children taking controller medication do so also on a frequent basis. Among the 10 states, Florida has the smallest proportion of children with frequent and high utilization of conroller medications (62%) compared to Texas with the highest proportion (90%).

While short-term medication utilization is much lower than the controller medication across all states, we find that there are groups of children who entirely rely on short-term medication. Among those children who do not use controller medication, the probability of a severe outcome, including emergency department encounter or hospitalization, is high. Utilization profiles (except one in Mississippi and one in Florida) with severe outcomes

also have no transtion from short-term medication to controller medication, indicating the importance of controller medication for persistent asthma. If asthma controller medication is present with high probability in the utilization profile, then it is also the follow-up event after a hospitalization or Emergency Department visit. The proportion of children in cluster with high prevalence of severe outcomes but no transition in controller medication varies across the states, from 4.7% in Texas to 22.9% in South Carolina.

While we only observe the prescriptions filled by a pharmacy and not those prescribed by a provider, we generally find that there is a higher probability of a controller medication than a short term medication from a physician office visit. This is not the same after a hospitalization or an Emergency Department visit, where there are states where short-term medication has a higher transition than controller medication. An important recommended guideline for care after a hospitalization or Emergency Department encounter is the follow up with a physisican office. For all states, the probability of such follow-ups is low, with a maximum of 0.38 in Tennessee but as low as zero for some utilization profiles in Alabama, Arkansas and South Carolina. In fact, there is a higher probability of yet another a hospitalization or Emergency Department encounter, with a probability as high as 0.58 in some profiles in Arkansas and Mississippi.

2.4.1 Limitation

One shortcoming of this study is reliance on claims data to infer utilization. First, the MAX files only include claims that have been submitted for reimbursement. Second, many Medicaid-eligible children have intermittent enrollment. Moreover, there will be a percentage of Medicaid-enrolled children who are undiagnosed due primarily to lack of healthcare access. Therefore, estimates on the healthcare utilization are likely to be biased, particularly for the Medicaid population, where certain subgroups have difficulty in maintaining

Medicaid coverage or are susceptible to particularly disparate utilization [20][21]. Moreover, Medicaid MAX files can have data quality issues, especially for states with large populations on managed care[22][23].

While our model and its estimation and selection methods are computationally attractive, they can be extended further by relaxing some of the underlying assumptions. First, we do not include the mean times until the first event and the mean times between the last event because they are biased estimates of complete lifetimes due to the censored nature of our data. Therefore, we are unable to completely determine the consistency with which patients visit providers or take medication. For instance, with unbiased estimates of the arrival to the first event it would be clear if a patient waits a long time between groups of consecutive visits or utilizes the system at a fairly homogeneous rate across the complete study time span.

Furthermore, in order to produce simple visualizations and minimize computational costs we assume the interarrival times to be exponentially distributed, conditional on the visit type. More importantly, it is likely that covariates including age, condition severity, comorbidities, enrollment status and access play a role in the frequency of the visits. However, this method does not capture the potential effects of these covariates on utilization.

2.4.2 Conclusion

Even though this study has several limitations, it has some important implications for health care providers and policy makers. Among the recommended care guidelines, we find that the highest level of adherence is with respect to asthma controller medication. In most profiles, children who take medication do so regularly. While there are children with persistent asthma who primarily rely on short-term medication, we also find multiple fold higher uti-

lization of asthma controller medication versus short-term medication unlike some existing studies. The lowest level of compliance is with follow-up physician office visits after an emergency room encounter or hospitalization. Since this finding is prevalent across all states, it is important to draw attention on the lack of compliance to this guideline at the national level. Finally, all states have a proportion of children who have uncontrolled asthma, meaning they do not take controller medication while they do have severe outcomes; the proportion varies significantly from one state to another with Texas having a very small proportion (below 5%) and South Carolina having the highest (higher than 20%).

CHAPTER 3

REGULARIZED OPTIMIZATION WITH SPATIAL COUPLING FOR ROBUST DECISION MAKING

In high-dimensional optimization problems where large number of decisions need to be made by optimizing a single objective, the resulting solution may exhibit high sensitivity to input parameter perturbations, which hinders reliable decision-making. This chapter introduces a regularized optimization approach to control the trade-off between optimality and sensitivity of the solution to optimization problems used to match supply and demand over a geographical area. The proposed regularization technique achieves some form of smoothing of the solution over the geographic area, motivated by the need of modeling intrinsic spatial dependencies between decision variables (called herein *spatial coupling*), thus resulting in a more realistic solution for applications. We demonstrate the wide applicability of the proposed approach for multiple optimization problems. We illustrate the proposed approach using a specific application in health care access measurement, in which a smooth solution that is robust to perturbations of model parameter leads to reliable decision-making. The experimental results show that the proposed approach can be used to find a smooth and robust solution while sacrificing its optimality at a minimum level.

3.1 Introduction

Spatial coupling is common in optimization models for optimal allocation of resources over a geographic area, for example, in transportation problems, facility location problems, evacuation planning, demand estimation, and measurement of access to fundamental services [24, 25, 26, 27, 28, 29, 30]. Spatial coupling arises due to the spatial structure intrinsic to the decision variables. For example, given a geographic region divided into small communities, one could be making decisions on how to assign sub-populations within the

communities to vaccination sites in the event of an emergency response, in such a way to minimize travel distance or congestion experienced, under a series of constraints, e.g., population assigned to a site shall not exceed total vaccine available and priority of children and pregnant women to the vaccine [31]. The decision variables are the proportion of sub-populations within each community to be assigned to each vaccination site. In a typical formulation, the decision variables are related to each other via supply and demand constraints. However, in applications, there are spatial dependencies beyond what the constraints can represent, that is, the decision made in one community will affect those made in the communities nearby (e.g., the first law of geography).

When making global or centralized decisions based on an optimization model over a geographic area, the output (optimal) decisions can vary significantly with small disturbances in the input parameters of the optimization model. The output decisions corresponding to nearby communities can change significantly because resources may shift from one community to another with slight changes of the input parameters in either the constraints or the objective, especially under limited resources. Moreover, this change may be cascaded to nearby communities. Thus, the optimal solution to the optimization problem will be unstable locally while maintaining optimality over the whole area. This is particularly evident in the context of high-dimensional optimization problems where a large number of decisions need to be made by optimizing a single objective function.

In statistical learning, controlling sensitivity of an estimator translates to controlling the bias-variance trade-off (e.g., see [32]). While unbiasedness of an estimator is a desirable statistical property, it may come with a price – high variability of the estimator with perturbations of the observed data. High variability in estimators might lead to unreliable decision making and prediction. Hence, it is often preferable to sacrifice bias to guarantee reliability. Finding a statistical model that balances biasedness and reliability is the objective of the bias-variance trade-off.

For an optimization problem, the bias and the variance in the bias-variance trade-off

correspond to the optimality of the objective function and sensitivity of the optimal solution, respectively. Similarly to statistical modeling, a solution to an optimization problem is not reliable in decision making if it is not robust to small changes in the input data. For example, in emergency responses, if a small change in the capacity available at a site will result in a significant change in how people are assigned to emergency sites, the "optimal" decision can be costly and challenging to implement.

Dealing with sensitivity to small disturbances in optimization is the subject of the well-established research field of robust optimization. One research stream in robust optimization is finding a solution that optimizes the worst case value of the objective function over an uncertainty set [33, 34, 35]. The second is finding a solution that optimizes a weighted sum of the first and the second moment of the objective function value given a probability distribution [36, 37]. Two limitations of both approaches are: 1. They require specification of an uncertainty set or distribution, which often is challenging to identify; and 2. The focus is on robustness of the optimal value rather than of the optimal solution whereas our focus is sensitivity of the optimal solution.

A common approach in statistical modeling and machine learning to controlling the bias-variance trade-off is by means of *regularization*. The concept of regularization is very general, from penalization for complexity in statistical modeling [38, 39, 40, 41] to smoothing of functional data [42]. The underlying idea is to consider additional information in the estimation objective function (e.g. likelihood function) to prevent from overfitting or from selecting complex models. Development of computationally efficient algorithms towards solving optimization problems arising in regularized statistical models has led to the emergence of the field of statistical optimization (see [41] and references therein). The methodology in this chapter is inspired by the advancements in this field however we consider regularization not of a statistical model but of an optimization model to deal with the sensitivity of the output decision. To the best of our knowledge, this is the first approach to addressing sensitivity of optimal decisions derived using optimization models,

by borrowing ideas from regularized statistical modeling.

In this chapter, we introduce a regularized optimization approach for optimization problems with spatial coupling to balance the trade-off between optimality and sensitivity, imposing some form of smoothing in the output solution. Smoothness will force the decision variables of nearby spatial locations to look "similar", thus preempting a drastic shift of resources from one community to another at time of small changes in the system. Similarity between decision variables or their transformation is defined depending on the decision outcomes that are of interest. To achieve smoothness of the output solution using regularization, a smoothness penalty is added to the objective function of the optimization model, which is similar to nonparametric smoothing of functional data [42], regularization in statistical optimization [41], and the stream of robust optimization methodology, which minimizes a weighted sum of the (expected) objective function value and the variance [36, 37]. The proposed approach applies a statistical technique to optimization, thus belongs to the less common direction of influence between statistics and optimization [43].

This chapter is structured as follows. We introduce the general regularized optimization problem in Section 2. We then illustrate it with well-established optimization settings in Section 3. We also provide a specific case study demonstrating its applicability and performance with respect to non-regularized version of the optimization problem in Section 4. We conclude with a summary of the proposed approach and future challenges to be considered in Section 5.

3.2 Regularized Optimization: General Approach

Although using regularization to control variability of optimal decision is a general approach, we will illustrate it in this paper on the class of optimization problems for matching resources (supply) and people (demand). Because the decision variables reflect transportation or resource allocation decisions for people living at spatially distributed demand locations, there is an intrinsic spatial coupling between the decisions of nearby demand

locations, reflecting similarities in barriers or preferences on how people are matched to supply sites. Thus, the regularization will not only address the sensitivity of the optimal solution but also reflect allocation of resources more evenly across neighboring communities, particularly important in allocation of public resources.

Let θ_{ij} denote the matching variable from demand location $i \in I$ to supply facility $j \in J$. Let Θ denote the $|I| \times |J|$ matrix whose (i, j) entry is θ_{ij}, and $\theta_{i\cdot}$ and $\theta_{\cdot j}$ denote the ith row and the jth column of Θ, respectively. Let G be a graph whose nodes are demand locations and for a demand location i, let $\delta(i)$ denote the neighbors of i in G. A general regularized optimization model is:

$$\text{(GP)} \ \min_{\Theta} \ F(\Theta) + \sum_{i \in I} \sum_{k \in \delta(i)} f_{ik}(\phi_i(\theta_{i\cdot}), \phi_k(\theta_{k\cdot})), \tag{3.1}$$

$$\text{s.t. } g_i(\theta_{i\cdot}) \leq 0 \text{ for } i \in I, \tag{3.2}$$

$$h_j(\theta_{\cdot j}) \leq 0 \text{ for } j \in J, \tag{3.3}$$

$$H(\Theta) \leq 0, \tag{3.4}$$

$$\Theta \geq 0$$

where F is a performance measure of matching Θ, ϕ_i is a real-valued function that represents a local performance of demand location i, f_{ik} is the objective term representing the disparity of performance at demand locations i and k (e.g., the 2-norm distance), (4.2) and (4.3) are local constraints for each demand location and supply facility, respectively, and (3.4) is a global constraint.

The regularization function $f_{ik}(\phi_i(\theta_{i\cdot}), \phi_k(\theta_{k\cdot}))$ has the role of forcing the solutions of the decision variables $\theta_{i\cdot}$ and $\theta_{k\cdot}$, or their transformations, $\phi_i(\theta_{i\cdot})$ and $\phi_k(\theta_{k\cdot})$, be "similar", where the similarity criterion is given by the function f_{ik}. This can be seen as an approach to smoothing the decision solutions or their transformations. For example, im-

posing smoothness under spatial coupling, the regularization function can be defined by

$$f_{ik}(\phi_i(\theta_{i\cdot}), \phi_k(\theta_{k\cdot})) = \lambda \|\phi_i(\theta_{i\cdot}) - \phi_k(\theta_{k\cdot})\|_2$$

Other norms can be considered, for example, L^1 norm or minimum, however the computational complexity and effort to solve the resulting optimization problems is greater than when using the L^2 norm [44].

The regularization parameter λ controls the trade-off between optimality and smoothness (and sensitivity as demonstrated in Section 4). The larger λ is, the smoother the decision solutions or their transformations are. A challenge in regularized optimization is finding the level of regularization specified by λ that best balances optimality and sensitivity. In Section 4, we illustrate a computational approach to identifying the regularization parameter λ using a modified bias-variance tradeoff approach where we find the balance between the "original" objective function (F in (4.1)) and the regularization function value (the double sum in (4.1)).

3.3 Regularized Optimization with Spatial Coupling: Applications

In this section, we introduce two illustrative examples to demonstrate the applicability of the regularization approach by smoothing: telecommunications network and evacuation planning. A third example, health care access measurement, is developed in more detail in the next section.

3.3.1 Telecommunications Network Design

The first example is an assignment problem arising in design of telecommunication networks [45, 46]. A graph $G = (N, A)$ representing a communication network is given, where N is the set of nodes and A is the set of arcs. When a demand for service from node i to node j is received, routes between the two nodes should be identified with sufficient

bandwidth, otherwise the demand is lost. The goal is to minimize the total volume of unmet requests.

The optimization problem is given as follows:

$$\begin{aligned}
&\min_{x,y} \sum_{i,j\in N} y_{ij} \\
&\text{s.t.} \sum_{i,j\in N} \sum_{p\in P_{ij}} \delta_{ap} x_p \leq u_a \text{ for } a \in A \\
&\qquad \sum_{p\in P_{ij}} x_p + y_{ij} = r_{ij} \text{ for } i,j \in N \\
&\qquad x \geq 0, y \geq 0,
\end{aligned}$$

where the variable y_{ij} denotes the lost demand from i to j and the variable x_p denotes the service volume on path p; P_{ij} denotes the set of paths in G from i to j; for a path p and an arc $a \in A$, δ_{ap} equals 1 if a belongs to p and 0 otherwise; u_a is the capacity of arc a; and r_{ij} denotes the demand of service from i to j.

In this application, the local performance of node i can be represented as $\phi_i(y_{i\cdot}) \triangleq \sum_{j\in N} y_{ij}$, that is, the total volume of unmet requests from node i. One can smoothen the local performance over the geographic area by adding penalty terms, for example, $||\phi_i(y_{i\cdot}) - \phi_j(y_{j\cdot})||_2^2/d_{ij}^2$ for $i,j \in N$, where d_{ij} denotes the geographical distance between i and j. Smoothing the local performance over the geographic area will prevent two nearby locations from having vastly different rates of lost service, which is, for example, important for identifying areas with poor communication performance.

3.3.2 Evacuation Planning

The goal of evacuation planning is to find a traffic diversion strategy over a geographic area in order to minimize the total evacuation time under some emergency conditions [47, 48]. It is defined on a graph $G = (N, A)$. The node set N is partitioned as $N = N_e \cup N_t \cup N_s$, where N_e, N_t, and N_s are the sets of evacuation, transit, and shelter nodes, respectively.

The optimization problem formulated in [47] can be written as follows:

$$\begin{aligned}
\min_{x} \quad & \sum_{i \in N_e} \sum_{j \in N_s} \sum_{p \in P_{ij}} c_p x_p \\
\text{s.t.} \quad & \sum_{i,j \in N} \sum_{p \in P_{ij}} \delta_{ap} x_p \leq u_a \text{ for } a \in A, \\
& \sum_{p \in P_{ij}} x_p = h_i \text{ for } i \in N_e, \\
& x \geq 0,
\end{aligned}$$

where the variable x_p represents the number of people evacuating through route p; c_p is the cost of route p (e.g., reflecting its distance); P_{ij}, δ_{ap}, and u_a are defined as in the previous example; and h_i denotes the number of people evacuating from node i.

The average evacuation distance from node i is $\phi_i(x) \triangleq \sum_{j \in N_s} \sum_{p \in P_{ij}} c_p x_p$ and one can smoothen the local performance by adding penalty terms, such as $||\phi_i(x) - \phi_j(x)||_1 / d_{ij}$ for $i, j \in N_e$, where d_{ij} denotes the geographical distance between i and j. Smoothing the average evacuation distance will result in a more equitable evacuation plan in which nearby communities do not have vastly different travel distances to evacuate.

3.4 Case Study: Health Care Access Measurement

In this section, we present another application of regularized optimization with spatial coupling, measuring spatial access to health care services. We introduce a regularized optimization model and demonstrate the impact of regularization on the smoothness and the sensitivity of solution, illustrating how the regularized model leads to more reliable decision-making in the application.

3.4.1 Optimization Model without Regularization

Spatial access refers to availability (provider-to-patient ratio) and accessibility (travel distance). It is a manifestation of the dynamics between a service provider network and needs for the service over a geographic area. Recently, linear programs (LPs) have been implemented to derive a matching between the supply and need, where access measures are formalized as linear functions of the assignment derived from the LP, such as the average travel distance to services at the community level [24]. The linear optimization model stratifies matching variables by financial access (e.g., insurance type) and other population characteristics related to access (e.g., health status, age).

The goal of the optimization model is to find a matching between those in need of health care and care providers satisfying the following properties: (1) minimizing the total distance traveled, (2) accounting for the preference of people not being willing to access providers with long wait times because of the large volume of patients, and (3) matching as many people in need as possible to providers, if not all need in the system can be satisfied. Other properties can be integrated in the optimization model as well. The model incorporates constraints on modes of transportation and limited service capacity for public insurance, among others.

For this illustrative example, we consider access to primary care for children in the state of Georgia. The model aims to quantify disparities in spatial access to pediatric primary care at different census tracts in Georgia. Providers' practice location addresses are obtained from the 2013 National Plan and Provider Enumeration System (NPPES). The 2009 MAX Medicaid claims data obtained from the Centers for Medicare and Medicaid Services are used to determine which providers have seen Medicaid patients. The patient population is aggregated at the census tract level, using the 2010 SF2 100% census data and the 2012 American Community Survey data to compute the number of children in each census tract along with information on household ownership of cars in order to estimate access to private transportation means. More details about the applied problem can be found in [24].

In the model to be presented in this paper, we assume that all children need the same level of care but we differentiate them by their insurance types (Medicaid or others), modes of transportation (owning a vehicle or not), and locations. In this context, I is the set of census tracts and J is the set of providers. The linear program for this assignment problem is:

$$\text{(AP)} \min_{\mathbf{x}_{\cdot\cdot},\mathbf{y}_{\cdot\cdot}} F(\mathbf{x}_{\cdot\cdot},\mathbf{y}_{\cdot\cdot}) \triangleq a_1 \sum_{i\in I}\sum_{j\in J} v d_{ij}(x_{ij}+y_{ij}) - a_2 \sum_{j\in J} u_j\Big(1-\frac{1}{\bar{c}_j}\sum_{i\in I} v(x_{ij}+y_{ij})\Big) \tag{3.5}$$

$$\text{s.t. } \sum_{i\in I}\sum_{j\in J}(x_{ij}+y_{ij}) \geq qQ, \tag{3.6}$$

$$v\sum_{i\in I}(x_{ij}+y_{ij}) \leq \bar{c}_j \text{ for } j\in J, \tag{3.7}$$

$$v\sum_{i\in I} x_{ij} \leq \bar{c}_j^{\text{med}} \text{ for } j\in J, \tag{3.8}$$

$$v\sum_{i\in I}(x_{ij}+y_{ij}) \geq \underline{c}_j \text{ for } j\in J, \tag{3.9}$$

$$\sum_{j} x_{ij} \leq q_i^{\text{med}}\cdot p_i \text{ for } i\in I, \tag{3.10}$$

$$\sum_{j} y_{ij} \leq (1-q_i^{\text{med}})\cdot p_i \text{ for } i\in I, \tag{3.11}$$

$$\sum_{j:d_{ij}\geq d_{\text{mob}}^{\max}} y_{ij} \leq m_i^{\text{oth}}\cdot(1-q_i^{\text{med}})\cdot p_i \text{ for } i\in I, \tag{3.12}$$

$$\sum_{j:d_{ij}\geq d_{\text{mob}}^{\max}} x_{ij} \leq m_i^{\text{med}}\cdot q_i^{\text{med}}\cdot p_i \text{ for } i\in I, \tag{3.13}$$

$$x_{ij}, y_{ij} = 0 \text{ for } i\in I, j\in J \text{ such that } d_{ij}\geq d^{\max}, \tag{3.14}$$

$$x_{ij}, y_{ij} \geq 0 \text{ for } i\in I, j\in J, \tag{3.15}$$

where

- Variables:

- x_{ij}: the number of Medicaid children in census tract $i \in I$ assigned to provider $j \in J$,
- y_{ij}: the number of non-Medicaid (e.g., privately-insured) children in census tract $i \in I$ to provider $j \in J$,

- Parameters:
  - v: the number of yearly visits required by a child,
  - d_{ij}: the distance between the centroid of census tract i and provider j,
  - $\bar{c}_j$: the maximum number of visits that can be accommodated by provider j,
  - u_j: a weight assigned to each provider to represent how sensitive children are for congestedness of provider j,
  - a_l for $l = 1, 2$: weights that balance the trade off between the terms in the objective function,
  - Q: the total population considered,
  - q: the minimum percentage of children to be assigned overall,
  - $\bar{c}_j^{\text{med}}$: the maximum number of Medicaid visits provider j can accommodate,
  - $\underline{c}_j$: the minimum number of visits that provider j needs to remain in practice,
  - q_i^{med}: the percentage of Medicaid children in census track i,
  - p_i: the number of people seeking service at census tract i,
  - $d_{\text{mob}}^{\text{max}}$: the maximum distance people without a vehicle are willing to travel for service,
  - m_i^{med}: the percentage of Medicaid children owning a vehicle at census tract i,
  - m_i^{oth}: the percentage of non-Medicaid children owning a vehicle at census tract i, and
  - d^{max}: the maximum distance any person would be willing to travel for service.

Interpretations of the objective function and each constraint are in order.

The first component of the objective function is the total distance traveled to receive services. In the second component, for a provider j, $\frac{1}{\bar{c}_j}\sum_{i\in I} v(x_{ij}+y_{ij})$ is the ratio of the assigned number of visits to the maximum number of visits provider j can accommodate. Therefore, the second component represents the provider preference contingent upon congestion level. The objective function is a weighted sum of the two components.

We maximize the number of matched people by using the constraint (3.6) and setting q as the maximum value for which the problem is feasible. The constraints (3.7) and (3.8) ensure that the total number of visits to a provider j do not exceed its capacity for both private and public insurance. The constraints (3.9) ensure that there is enough demand for each provider to remain in practice. (3.10) and (3.11) ensure that the assignment of children from census tract i to all providers does not exceed the population in need at i, for the two insurance types. The constraints (3.12) and (3.13) represent the access barriers due to the ownership of a vehicle. (3.14) impose the maximum travel distance for service and lastly, the constraints (3.15) ensure the variables are nonnegative.

The optimization model provides an optimal assignment over the whole state of Georgia that matches children in each census tract to providers' locations under the constraints. Based on the assignment, healthcare access of census tracts can be compared by calculating the average distance that children at each census tract need to travel to reach a health care provider. In the following, z_i and w_i are the access measure for children enrolled in Medicaid and those with other forms of financial access for census tract i, respectively, a weighted average of the traveling distance for each census tract i and provider j pairs:

$$z_i = d^{max} + \frac{1}{q_i^{\text{med}} \cdot p_i}\sum_{j\in J}(d_{ij}-d^{max})x_{ij} \text{ for } i \in I \text{ and} \tag{3.16}$$

$$w_i = d^{max} + \frac{1}{(1-q_i^{\text{med}}) \cdot p_i}\sum_{j\in J}(d_{ij}-d^{max})y_{ij} \text{ for } i \in I. \tag{3.17}$$

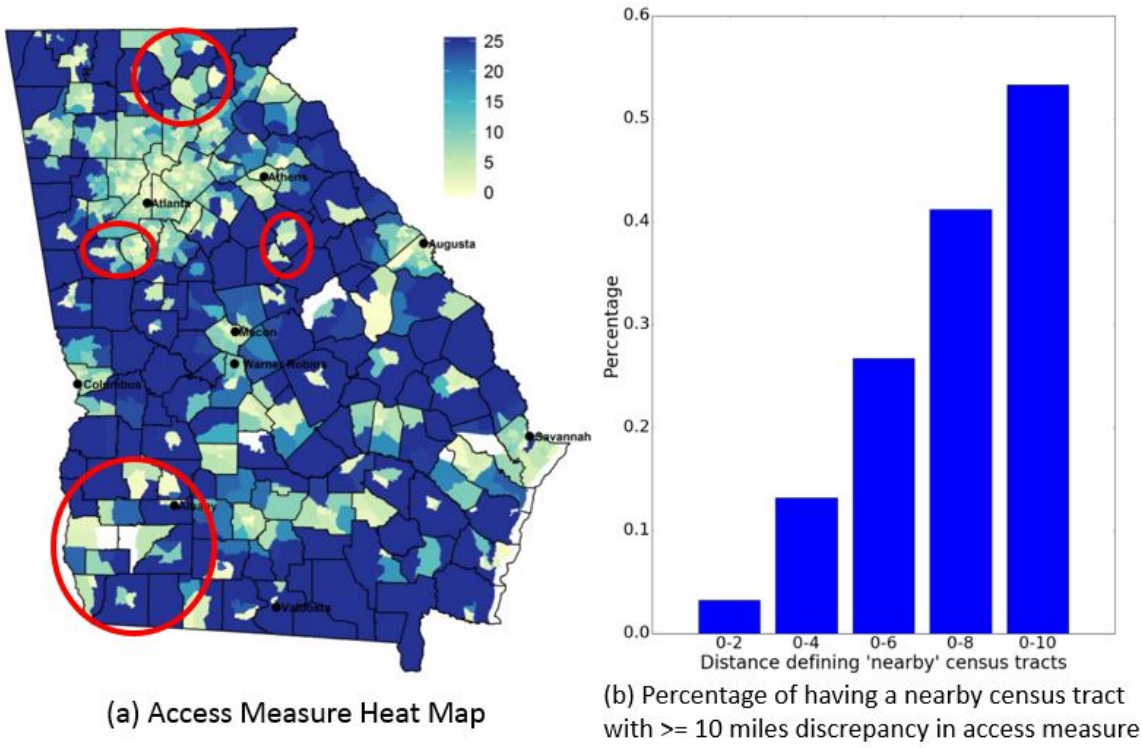
Note that for those who are not assigned to a provider, the maximum travel distance ($d^{\max}$) is assumed, and thus w_i and z_i range from 0 to $d^{\max}$.

These access measures are used to identify areas that have low access. Limited resources (e.g., a new facility or providers supported by public agency) are to be deployed targeting those areas.

The problem (AP) is a linear optimization problem, with a large number of variables for each pair of census tract and provider, and each insurance type. For example, the state of Georgia has 1955 census tracts and 3157 provider locations, and we consider two types of financial access, public and private. The total number of variables that are allowed to be nonzero (i.e., after accounting for the constraint (3.14)) is approximately 2.1 million.

Figure 1(a) shows the average travel distance in miles for all census tracts in Georgia, ranging from 0 to 25, computed by (AP). In lighter areas, mostly major cities or areas close to major cities, children do not need to travel more than 5 miles to reach a primary care provider whereas children in the darker census tracts, which are mostly rural areas, have to travel near or at least 25 miles. Note that children living in some nearby census tracts, circled in red, exhibit vastly different access measures under the model (AP). In those regions, there are pairs of census tracts next to each other with a travel distance difference of nearly 20 miles. Figure 1(b) shows what percentage of census tracts have another census tract nearby whose travel distances differ more than 10 miles, for different distances between census tracts. More than half of the census tracts have at least one census tract within 10 miles such that their access measures differ by at least 10 miles. These large differences between nearby census tracts result from the nature of the problem formulation, minimizing the total travel distance over the entire state with limited resources. However, intuitively, people living in nearby locations should experience similar level of access to health care providers.

In addition, the resulting access measures can be sensitive to small perturbations of the parameters in the optimization model. Because the parameters estimated from data in-



(a) Access Measure Heat Map

(b) Percentage of having a nearby census tract with >= 10 miles discrepancy in access measure

Figure 3.1: (a) Heat map of access measure in average traveling distances. (b) Percentage of census tracts with at least one neighboring census tracts differ more than 10 miles in access measure, broken down by the distance between the centroids of the census tract pair.

volve uncertainty, the sensitivity of access measure hinders reliable decision-making. For illustration, we generated 100 sets of parameters varying within small ranges, and computed the corresponding access measures of each census tract for each parameter set. The parameter sets were obtained by multiplying the maximum capacity $\bar{c}_j$ of each provider j by a constant, independently sampled from the uniform distribution on $[0.8, 1.2]$. Such perturbations in the capacity of a provider can be due to increasing and/or decreasing of personnel, overtime or days off of providers, inaccurate estimation of capacity among others. Since the random perturbation has zero mean, the total provider capacity does not change in expectation. Figure 2 shows the range of access measure of each census tract, that is, the difference between the maximum and the minimum access measures over the 100 runs. The darker the color, the more unstable the measure is. Census tracts in gray are the ones whose access measure varied more than 10 miles. 85 out of 1955 census tracts

changed more than 5 miles in access measure, and 22 census tracts changed more than 10 miles. Considering the access measure ranges from 0 to 25 miles, this implies that some census tracts may have either high or low access, when parameters are perturbed within a realistically small range. Thus, decision-making to allocate limited resources using the optimization model (AP) is vulnerable to small perturbations of the model parameters.

This phenomenon may arise in the more general context of high-dimensional resource allocation problems. A small perturbation in the input parameters may cause a dramatic shift in assignment. A solution that is susceptible to uncertainty in input parameters makes it hard to infer sensible recommendations to decision makers.

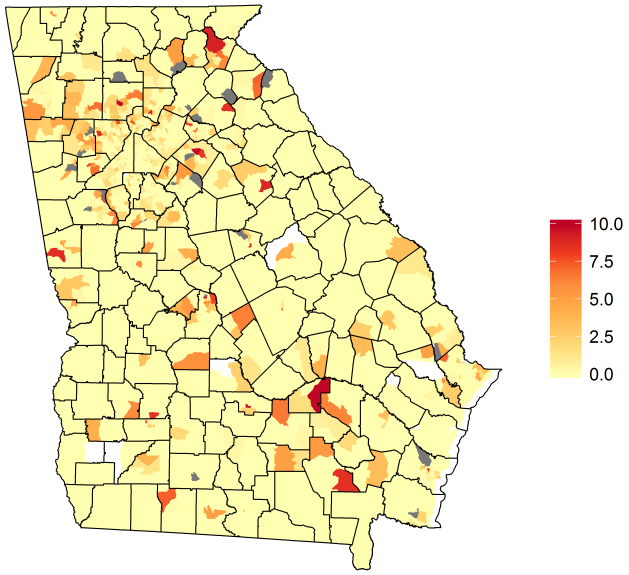


Figure 3.2: Range of access measure for each census tract from the 100 runs with slightly different provider capacity.

3.4.2 Regularized Formulation

Using the motivating application in the previous sub-section, we have demonstrated the non-smoothness and sensitivity issues in high-dimensional optimization problems with large number of decisions made by optimizing a single objective function. In this section,

we introduce a regularized formulation of (AP) and demonstrate smoothness and reduced sensitivity of the access measure obtained by the new model. Motivated by the intuition that children living in neighboring areas should experience similar level of health care access, we add penalty terms that control smoothness of the access measure. For each pair of census tracts within a specified distance, the objective function has an additional term, the squared difference between their average travel distances divided by the distance between the two census tracts, thus giving a penalty for difference of their access measures, inversely proportional to their distance. The closer two census tracts are, the more penalty for a unit difference of average travel distances. The new formulation is as follows:

$$\text{(RAP)} \min_{\mathbf{x}_{\cdot\cdot},\mathbf{y}_{\cdot\cdot},\mathbf{z}_{\cdot\cdot},\mathbf{w}_{\cdot}} (1-\lambda)F(\mathbf{x}_{\cdot\cdot},\mathbf{y}_{\cdot\cdot}) + \lambda\left(a_3\sum_{i\in I}\sum_{k\in S:d_{ik}\le d_{\text{pen}}^{\max}}\frac{1}{d_{ik}}(z_i-z_k)^2 + a_4\sum_{i\in I}\sum_{k\in S:d_{ik}\le d_{\text{pen}}^{\max}}\frac{1}{d_{ik}}(w_i-w_k)^2\right)$$

s.t. constraints (3.6) through (3.17),

where

- d_{ik}: the distance between the centroids of census tracts i and k,
- $d_{\text{pen}}^{\max}$: the maximum distance we penalize for difference of access measure (here we set it to be 10 miles),
- a_3, a_4 : weights that balance the trade-off between the terms in the objective function, and
- λ: regularization parameter which controls the degree of penalization.

The additional components in the objective have the role of smoothing the access measure. The regularization parameter λ dictates the amount of penalty being added to the

objective function. The closer to one λ is, the smoother the resulting access measure is on the spatial domain.

In order to determine a value of λ, we solved the (RAP) for varying λ values from 0 to 1 and plotted the values of $F(\mathbf{x}_{..}, \mathbf{y}_{..})$ (the "original" objective function) and the values of the regularization term (the additional objective terms) for the obtained solutions, in Figure 3. We observe that for small values of λ, say, from 0 to 0.2, the rate of change in F value is minimum while the value of the regularization term changes drastically. This is an indication that only a small amount of optimality of the original objective function is required to be sacrificed to significantly increase the level of smoothness. As λ further increases, we start to observe over-regularization, where the resulting solution vastly deviate from optimality to make room for negligible improvement on the regularized term. Therefore, the optimal choice of λ is where two curves are both approximately flat. We choose $\lambda = 0.55$ for the following experiments.

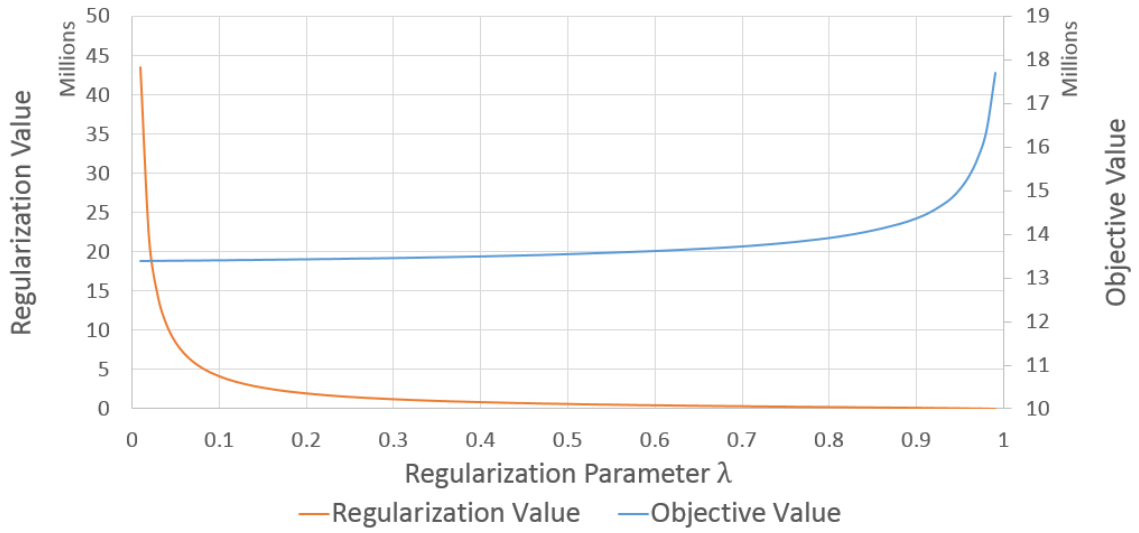


Figure 3.3: Trade-off between the objective function and regularization function values for varying regularization parameter λ

Figure 4(a) shows the average travel distance in miles for children in Georgia to reach a primary care provider, ranging from 0 to 25, computed by (RAP). Comparing with the result from (AP) (Figure 1(a)), the regularized model's output is smoother on the spatial

domain. Comparing Figure 4(a) with Figure 1(a) in more detail, regions circled in red, such as southwest of Atlanta, northeast of Atlanta, south of Athens, and census tracts close to Albany have access measures that are not drastically different from its nearby regions. Figure 4(b) shows the range of access measure of each census tract computed by (RAP) using the same 100 sets of perturbed parameters used for Figure 2. For the regularized model, only two census tracts had more than 5 miles of variation in their access measures, compared to 85 census tracts for the original formulation (AP).

In sum, Figure 3 shows that smoothness of access measure can be achieved with a minimal sacrifice of optimality of the original objective function and also provides a way to determine a proper value for the regularization parameter. Figure 4 demonstrates that the smoothing reduced sensitivity of access measure for perturbations of input parameters. In the context of access measure application, smoothing makes the resulting access measure more realistic and consistent with intuition and the reduced sensitivity enables decision makers to identify low access regions more reliably.

Another uncertain parameter in the access model is the number of required visits per year per child (v). This parameter varies depending on the recommended guidelines and/or demand for care. Unlike the capacity of an individual provider, the number of visits affects all terms in the objective function and all of the provider constraints, and thus, any change would impact the dynamic of the whole matching. We increased the parameter v by 0.01 from 2.80 to 3.20, solved (AP) and (RAP) for each v value, and computed the average travel distance of each census tract. Figure 5 shows the distribution of the change of the access measure of each census tract for each increment of v, as a box plot (the upper plot is for (RAP) and the lower one is (AP)). For example, the box at $v = 3.00$ shows the distribution of the change in the access measure of each census tract, where v changes from 2.99 to 3.00. We find that most of the change in (RAP) are controlled within plus or minus 2 miles, whereas the change in (AP) is unstable, even for a slight change (by 0.01) in the v value.

Similar results on the smoothness and sensitivity of outcome measures of the assign-

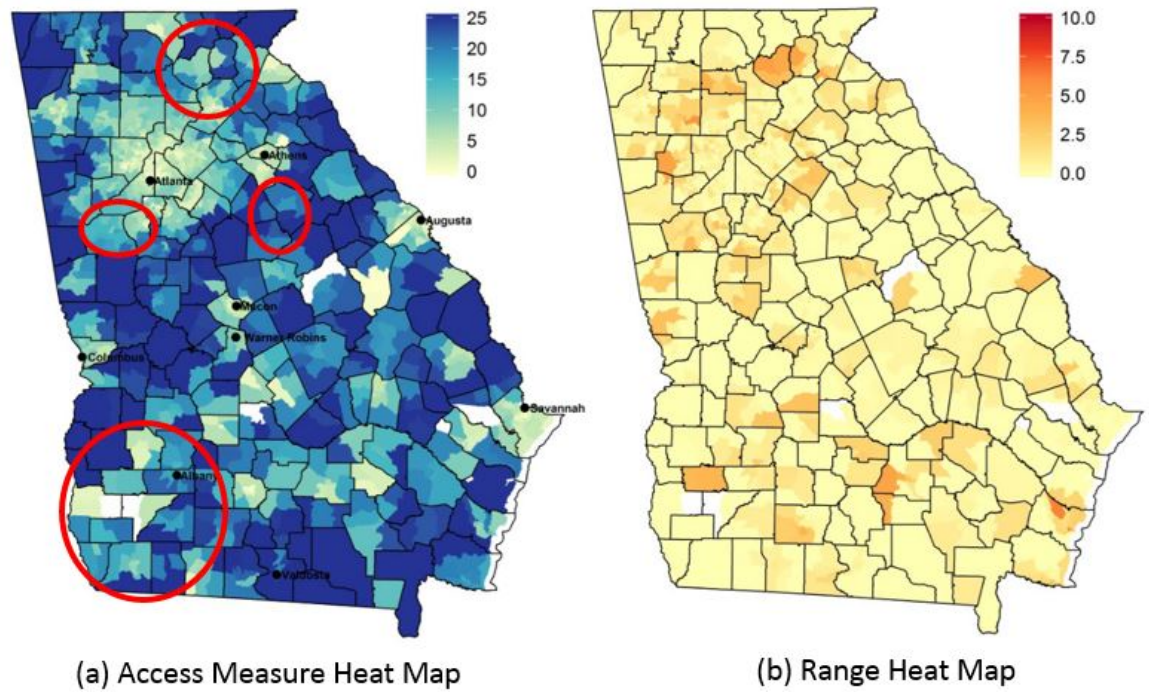


Figure 3.4: (a) Heat map of access measure in average traveling distances calculated from model RAP. (b)Range of access measure for each census tract from the 100 runs with slightly different provider capacity calculated from model RAP.

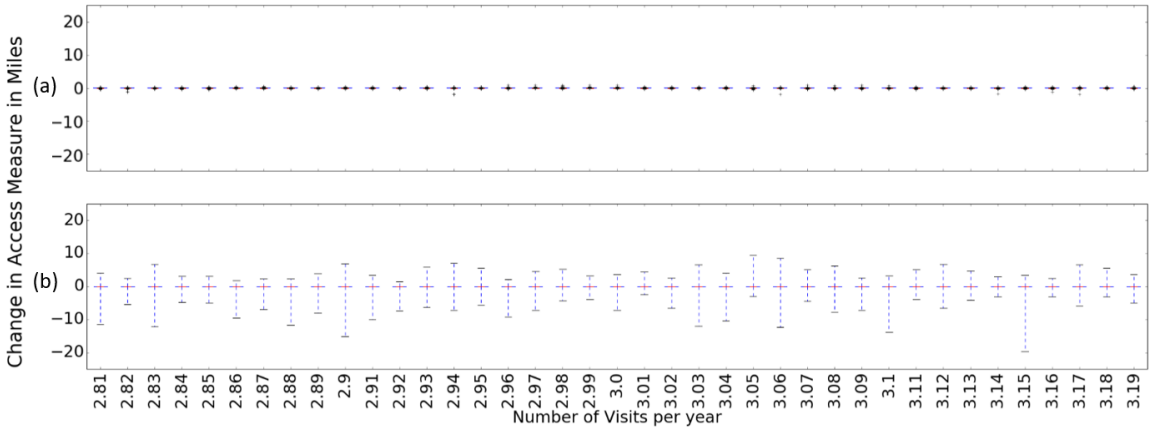


Figure 3.5: Box plots in change of access measure during each 0.01 increase in number of visits per year per child, calculated from regularized formulation (a) and original formulation (b).

ments can be derived for other model parameters in the motivating optimization problem. Overall, this application demonstrates the advantages of the proposed approach for both smoothness and robustness of the decision-making.

3.5 Discussion

This paper introduces a novel approach to making decisions from optimization models, incorporating local spatial dependence information while pursuing global robustness to small perturbations in the model parameters. The approach is inspired from established methodology for controlling the bias-variance trade-off. While the underlying idea is simple, it can be effective in achieving more meaningful decision making. The approach is also general, with application to multiple optimization problems as provided in this paper.

Centralized decision making without accounting for intrinsic dependencies could result in inequitable and ineffective implementation of the optimal decisions. This is particularly relevant for high-dimensional optimization models where a large number of decisions need to be made using a single objective. The constraints in the optimization models have the role of capturing a more realistic decision solution however if spatial or some other form of coupling is present, the solution needs to incorporate this knowledge. This aspect has been the topic of extensive research in statistical modeling, particularly in nonparametric regression [42] and in functional data [49], where the idea of borrowing information across dependent data leads to estimators that achieve a better fit in the sense of the bias-variance trade-off.

Borrowing information across spatially-dependent decisions derived using optimization models under spatial coupling results in a decision solution that can be more meaningful/realistic, as demonstrated in the applications provided in this paper. Accounting for spatial coupling is particularly relevant for decision making when the decisions are for highly granular geographic areas, for example, small communities or neighborhoods, because the spatial dependence between decisions is stronger.

When measuring healthcare access at the census tract level, a granular geographic division, communities in close proximity are expected to have similar access, according to the first law of geography. By not accounting for spatial coupling, the optimal solution to the optimization problem used to measure access can result in access estimates that are different by 5-10 miles for neighboring communities. If interventions are to be targeted based on such estimates, resources could be allocated inequitably.

The decisions made using regularized optimization are not only more realistic but can also be robust to small perturbations of the model parameters if the regularization penalty controls the sensitivity of the solution or functions of the solution. For example, the penalty considered in the motivating application can be interpreted as a measure of the variability in the outcome function of the decision, the access measure. The experiments provided in the motivating application indeed demonstrate robustness to variations in both global and local model parameters.

One potential limitation of the proposed approach is its higher computational complexity over the non-regularized version of the optimization problem, particularly if the penalty incorporates both L_1 and L_2 regularization. A second limitation is the specification of the regularization penalty, which needs to incorporate knowledge about the dependency in the outcome functions of the solution.

CHAPTER 4

VARIABLE PARTITIONING FOR DISTRIBUTED OPTIMIZATION

In this section, we introduce an approach for partitioning decision variables while decomposing a large-scale optimization problem for the best performance of distributed solution methods. Solving a large-scale optimization problem sequentially can be computationally challenging. One classic approach to address this computational challenge is to decompose the problem into smaller sub-problems and solve them in a distributed fashion. However, as we show in this paper, the decomposition of variables to form the set of sub-problems is key in reducing the computational effort when the optimization formulation involves complex constraints. To introduce and illustrate the approach to variable partitioning proposed in this paper, we focus on one of the most popular distributed optimization methods, dual decomposition and distributed sub-gradient methods. Based on a theoretical guarantee on its convergence rate, we explain that a partition of variables can critically affect the speed of convergence and highlight the importance of the number of dualized constraints. We consider various partitioning methods, ones that are based on given domain knowledge and others based on clustering algorithms. In particular, we introduce a novel partitioning approach that focuses on reducing the number of dualized constraints by utilizing a community detection algorithm from physics literature. Empirical experiments using a real application show that the proposed method significantly accelerates the convergence of the distributed sub-gradient method. The performance of the proposed method improves as the size of the problem increases, and as each constraint involves more variables.

4.1 Introduction

Solving large-scale optimization problems using one computer core and sequential computing can be computationally challenging due to the data storage and retrieval, and due to

the computational load and memory usage for obtaining an optimal solution. Distributed computing is a popular framework for tackling the computational complexity of large-scale optimization [50, 38, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60]. Distributed computing for optimization problems involves two computational considerations: the decomposition into smaller sub-problems in a way that each sub-problem can be stored and solved in a single machine, and the derivation of a solution for the original optimization problem by (iteratively) solving the sub-problems [38, 51, 53, 54, 55, 57, 58, 59, 60]. Applications of distributed optimization arise in various emerging areas, such as resource allocation over large-scale networks [53, 54, 60], aircraft coordination [52, 55], and estimation problem in sensor networks [61].

One classic approach to the decomposition of an optimization problem is the *dual decomposition*. It decomposes an optimization problem into smaller sub-problems by relaxing some of the constraints. Then, the resulting Lagrangian dual is often solved by a distributed sub-gradient algorithm [54, 55, 58, 59, 60]. Another decomposition technique is introducing copy variables and the alternating direction method of multipliers (ADMM) is commonly used in the distributed algorithm [38]. Dantzig-Wolfe (DW) decomposition is another popular method for large-scale optimization. However, each sub-problem in the DW decomposition is not a part of the original optimization problem. Solutions of the sub-problems in each iteration do not form a solution for the original optimization problem, but find nonbasic variables whose reduced costs are negative (in case of minimization) at the current basis. In this paper, we consider a methodology that decomposes the original optimization problem into sub-problems whose solutions form an approximate solution for the original optimization problem.

Most of the existing research has focused on developing either a new decomposition technique or a novel distributed algorithm. However, regardless of a decomposition technique (e.g., dual decomposition) or a distributed solution algorithm (e.g., sub-gradient method), partitioning the decision variables across sub-problems remains one of the key

challenges in distributed optimization. For instance, for a network optimization problem over a graph, should we define a sub-problem for each node or a group of nodes? If a sub-problem may contain multiple nodes, how should we assign nodes to sub-problems? Research discussing such aspects in distributed optimization is limited. In [62], a distributed block splitting algorithm based on graph projection splitting was introduced for decomposing and solving large-scale problems in parallel in which the objective function is separable by blocks of variables. However, the same paper pointed out that, in practice, it was not obvious which subset of variables should be processed together versus on separate machines. There have been other efforts to determine how a complex system should be partitioned to achieve faster convergence, for example, the spectral clustering technique in [63] and the simultaneous partitioning and coordination strategy in [64]. However, these works focused on specific applications with relatively small numbers of variables (a few hundreds or less).

The computational approach in this paper is motivated by the observation that a decomposition of an optimization problem (in other words, a partition of decision variables) critically affects the computational performance of distributed optimization algorithms. For illustration, we used one of the most common approaches in distributed optimization, dual decomposition and distributed sub-gradient method. Sub-gradient methods have been shown to converge to optimality as long as the resulting Lagrangian dual satisfies strong duality, regardless of which constraints are dualized or how decision variables are partitioned [65]. However, our empirical analysis shows that the convergence may be extremely slow, potentially not reaching convergence even after a large number of iterations (e.g., ten thousands), especially when there is a large number of highly complex constraints.

In this paper, we provide a theoretical explanation of why a partition of decision variables can affect the convergence of distributed sub-gradient methods. In relation to the theoretical upper bound for the convergence rate of sub-gradient methods established in literature [66, 67], we explain the importance of minimizing the number of dualized con-

straints to achieve faster convergence of distributed sub-gradient methods. The intuition is that the more "similar" the Lagrangian relaxation and the original problems are, the more desirable it is for the empirical performance of sub-gradient methods, yielding faster convergence results.

A key contribution of this paper consists of the novel methods to find a partition of decision variables for dual decomposition and to solve the resulting sub-problems with the blocks of variables. Our focus is dualizing as fewer numbers of constraints as possible while decomposing a large-scale optimization problem. We construct a graph representing the relationship between variables given by constraints and apply clustering algorithms to the graph to identify subsets of variables to be grouped together. In particular, we illustrate with a novel method using a *community detection* algorithm from the physics literature [68, 69]. The goal of community detection is to identify community structures within a network, in other words, to find groups of nodes in such a way the connections within each group are dense while there is little connectivity between the groups. Roughly speaking, we use community detection to group decision variables that tend to appear in constraints together so that the number of constraints that involve variables over multiple sub-problems (thus, need to be dualized) is minimized. Community detection has been applied to the Internet, citation networks, social networks among others [70]. In a recent work [71], community detection was used as a subroutine of an algorithm to find a Lagrangian relaxation for mixed integer programs with a tighter dual bound. On the other hand, our paper focuses on addressing the variable partitioning issue in the context of distributed optimization.

The proposed approach is general and applicable to various problem classes, but for illustration purpose, we present the proposed method applied to transportation problems as follows. First, we construct a graph whose nodes represent demand locations. Two nodes are connected if there is a constraint involving the two demand locations (e.g., a supply location can serve the two demand locations and there is a capacity limit constraint at the supply location). Each edge is weighted by the number of constraints involving the two

demand locations. Then, we apply a variable partitioning,for example, the community detection, to the graph to find a partition of demand locations. The community detection algorithm identifies communities of demand locations such that demand locations in the same community are densely 'connected' by the constraints but those in different communities are sparsely 'connected' by the constraints. Then, we decompose the original optimization problem into blocks of demand locations given by the community detection, thus reducing the number of dualized constraints by utilizing the community structure. Our empirical illustration in Section 4.4 shows that the method introduced in this paper significantly accelerates the convergence of distributed sub-gradient methods.

This paper is organized as follows. First, we review dual decomposition and sub-gradient methods in Section 4.2. In this section, we also analyze why decomposition is important for the performance of sub-gradient methods. In Section 4.3, we introduce several partitioning methods including the new approach using community detection. We illustrate its empirical performance for a real application in Section 4.4 and conclude in Section 4.5.

4.2 Dual Decomposition and Sub-gradient Method

Dual decomposition is a common technique for decomposing a large-scale optimization problem into smaller sub-problems [65, 55, 59]. Given a partition of decision variables, constraints that are over multiple groups of variables are relaxed and added to the objective function as penalty terms for violation, so that the Lagrangian relaxation is decomposable into smaller sub-problems. In this section, we first review the dual decomposition technique, followed by a distributed sub-gradient algorithm. We also analyze its convergence rate established in the existing literature and discuss why the convergence may be slow.

4.2.1 Transportation Problem

The transportation problem is a general class of problems, in which commodities are transported from a set of sources to a set of destinations. Let x_{ij} denote the matching variable from demand location $i \in I$ to supply location $j \in J$. Let X denote the $|I| \times |J|$ matrix whose (i, j) entry is x_{ij} and X_i denotes the ith row. The general optimization model is given as follows.

$$\text{(GP)} \min_{X} \sum_{i \in I} \sum_{j \in J} w_{ij} x_{ij} \tag{4.1}$$

$$\text{s.t.} \sum_{j \in J_i} x_{ij} \geq m_i \text{ for } i \in I, \tag{4.2}$$

$$\sum_{i \in I_j} x_{ij} \leq s_j \text{ for } j \in J, \tag{4.3}$$

$$X \geq 0, \tag{4.4}$$

where m_i is the minimum demand that needs to be satisfied at each demand location $i \in I$, s_j is the maximum capacity at each supply location $j \in J$, w_{ij} is the cost associated with demand location i getting one unit of goods from supply location j, J_i is the set of supply locations that can serve demand location i, and I_j is the set of demand locations that can be served by supply location j. In real applications where there is a large number of demand and supply locations, it is often assumed that each demand location can only be served by a subset of supply locations. For instance, in logistics, suppliers may have access only to a few demand locations due to regions of operations, or that some demand locations are simply too far away. In this paper, we consider only continuous decision variables. For example, x_{ij} may be a number of service hours assigned to demand location i from supply location j.

4.2.2 Dual Decomposition and Distributed Sub-gradient Method

In this section we review distributed sub-gradient method for (GP) with a straightforward partition of decision variables, a sub-problem for each demand location i. In order for (GP) to be decomposed for each i, all of the supply constraints (4.3) are relaxed and appended as penalties for their violation to the objective function. Let $\lambda_j \geq 0$ be the dual variable for each constraint in (4.3). The local sub-problem is written as follows:

$$\begin{aligned} (\text{LR}_i) \ \min_{X_i} \ L_i(X_i, \Lambda) &= \sum_{j \in J_i} (w_{ij} x_{ij} + \lambda_j x_{ij}) \\ \text{s.t.} \ \sum_{j \in J_i} x_{ij} &\geq m_i, \\ X_i &\geq 0. \end{aligned}$$

A distributed sub-gradient algorithm for solving the Lagrangian dual is given as follows.

Distributed Sub-gradient Algorithm

1. Choose a starting point Λ^1. Let $t := 1$ (first iteration).

2. Solve the local optimization problem (LR_i) with $\Lambda = \Lambda^t$ for each demand location $i \in I$ to obtain X_i^t.

3. If a given stopping criterion is satisfied, stop. Otherwise, $t := t + 1$, update the multipliers as below, and go to Step 2:

$$\lambda_j^{t+1} = \max\{\lambda_j^t + \alpha_t (\sum_{i \in I_j} x_{ij}^t - s_j), 0\} \text{ for } j \in J. \tag{4.5}$$

It is well-known that if the step-size $\{\alpha_t\}_{t=1}^{\infty}$ satisfies

$$\sum_{t=1}^{\infty} \alpha_t = \infty \text{ and } \sum_{t=1}^{\infty} \alpha_t^2 < \infty, \tag{4.6}$$

then the value of $g(\Lambda^t)$ converges to the optimal objective function value of (GP) (e.g., see [65]). Moreover, the running average of the primal iterates X^t becomes asymptotically optimal for (GP) as t goes to infinity [72, 58].

4.2.3 Analyzing Convergence Rate of Sub-gradient Method

Convergence rates of sub-gradient methods have been established under various settings [66, 67]. We first review a convergence rate result of the sub-gradient method with the step-size given in (4.6) and discuss its slow convergence and our proposed approach to address it.

Since sub-gradient methods do not improve monotonically, it is common to keep track of the best solution up to the current iteration. Let g denote the objective function of the Lagrangial dual and let Λ^t_{best} denote the solution having the lowest g value at the end of iteration t. Let R be an upper bound on the distance between the initial dual solution and the set of optimal dual solutions, i.e., $||\Lambda^1 - \Lambda^\star||_2 \leq R$. Also, let G be an upper bound on the norm of the sub-gradients computed by the algorithm, i.e., $||h^t||_2 \leq G$, where $h^t \in \mathbb{R}^{|J|}$ and $h^t_j = \sum_{i \in I_j} x^t_{ij} - s_j$ for $j \in J$. From [66], we have the following upper bound on the optimality gap, which goes to zero as t goes to infinity:

$$g(\Lambda^t_{\text{best}}) - g(\Lambda^\star) \leq \frac{R^2 + G^2||\alpha||_2^2}{2\sum_{k=1}^t \alpha_k}. \tag{4.7}$$

However, depending on the value of its numerator, the upper bound may converge to zero so slowly that it does not approach zero even at a large value of t (e.g., hundreds of thousands). See Figure 4.1 illustrating the significant difference in the convergence of the upper bound depending on the value of the numerator where $\alpha_t = \frac{1}{t}$. More importantly, the optimality gap itself may not approach zero even after a large number of iterations under the dual decomposition for each demand location. We emphasize that the slow convergence of the theoretical upper bound applies to any optimization problem, not limited to transportation

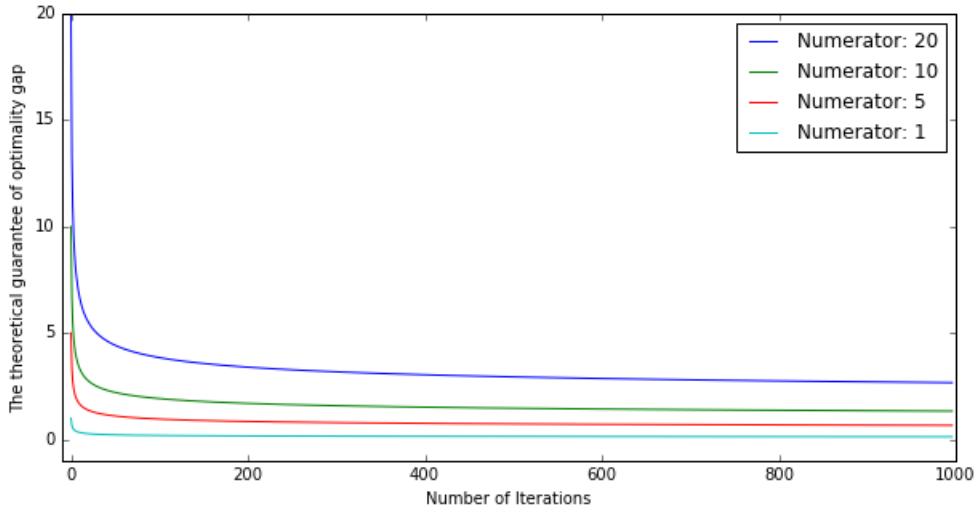


Figure 4.1: Convergence of the theoretical guarantee of optimality gap with different numerator values and $\alpha_t = \frac{1}{t}$.

problems used for illustration in this paper.

Next we discuss possible ways to speed up the convergence of the upper bound. The upper bound contains the step size $\{\alpha_t\}$, an upper bound R on the distance between the initial dual solution and the optimal dual solution set, and an upper bound G on the magnitude of the sub-gradients. Adjusting the step size is an easy choice for accelerating sub-gradient methods, but it is specific to each application and purely empirical. Another important component that governs the behavior of the upper bound is *the number of dualized constraints*, because it equals the dimension of dual parameter vector Λ and the dimension of the sub-gradients. Therefore, the number of dualized constraints is closely related to the magnitude of R and G, and thus, directly affects the convergence of the upper bound.

In addition, note that each component of the sub-gradient at a primal solution X is the violation of the corresponding dualized constraint at X. Thus, the fewer dualized constraints are violated (and the smaller magnitude the violations are), the lower the magnitude of the sub-gradient is. This again emphasizes the importance of the number of dualized constraints. Dualizing more constraints leads to a more relaxed feasible region of

the resulting Lagrangian relaxation. Then, the primal iterations obtained while running the sub-gradient method have more "room" to deviate from the original feasible region, thus allowing violations of more dualized constraints and also larger violations, which leads to higher magnitudes of the sub-gradients, and thus, a higher G value and slower convergence.

For the transportation problem and the dual decomposition introduced in the previous section, the intuition provided above is interpreted as follows. The Lagrangian relaxation of (GP) is obtained by dualizing all of the supply constraints (4.3), so the resulting relaxation differs significantly from the original problem. The sub-problem (LR_i) for demand location i is simply matching the demand of i to accessible supply locations where the values of the dual multipliers make the location i prefer some supply locations than others. Thus, the competition among demand locations for limited resources is only indirectly reflected via the dual multipliers. In other words, the level of decomposition is so fine that each sub-problem (LR_i) loses an important aspect of the original problem, which makes the overall convergence slow.

A partition of decision variables critically affects the computational performance of the sub-gradient method, and the goal of this paper is to develop a novel method to find a decomposition that speeds up distributed algorithms. In the context of dual decomposition, we aim at dualizing as fewer numbers of constraints as possible while taking the computational advantage of distributed computing. Consider decomposing (GP) into a given number of sub-problems by partitioning demand locations. The demand constraints (4.2) are decomposable by demand locations, but the supply constraints (4.3) are not. Given a partition of demand locations, those supply constraints involving demand locations in multiple groups need to be dualized in order for the remaining constraints to be decomposable. Herein we define two demand locations *connected* if and only if there exists a supply location that can serve both of the demand locations, i.e., they appear together in the capacity constraint of the supplier. Then, finding a decomposition with a minimal number of dualized constraints translates into finding a partition in which demand locations in the

same group are closely connected and those from different groups are loosely connected. We expand on this idea in the next section after discussing two other partitioning methods which are based on external knowledge.

4.3 Partitioning Methods and Block Dual Decomposition

In this section, we introduce a novel framework for distributed optimization, consisting of two steps:

• *Step 1*: Variable partitioning; and

• *Step 2*: Block dual decomposition.

The proposed approach uses the structure of constraints in order to speed up the convergence of distributed sub-gradient methods. We illustrate the approach using the transportation problem, but we also introduce a general version of the framework in Section 4.3.3.

4.3.1 Step 1: Variable Partitioning

In this section, we discuss three approaches to partitioning decision variables. The three approaches apply generally to networks with an established structure but we will specifically introduce the approaches in the context of geographically or spatially structured networks, such as in transportation problems. The first approach assumes that we have prior knowledge about grouping of the variables, for example, grouping the variables according to an established division of the geographic space. The second approach is employing a clustering algorithm using information about the similarity among the variables, for example, geographic distance between the regions corresponding to the variables. Lastly, we introduce a partitioning method using community detection, which utilizes the structure of the optimization problem itself.

The output of these approaches consists of blocks of variables; those variables assigned to different block are assumed to be de-coupled while those within the same block are assumed to be coupled within the distributed optimization problem.

Prior Knowledge.

In many applications, particularly in transportation problems, there may already exist some prior knowledge that suggests how the decision variables can be partitioned into different blocks. In cases where variables correspond to geographic locations, a grouping of the variables can be based on geographical sub-areas, such as county, census tracts, health districts, etc.

Clustering.

Clustering algorithms such as k-means can be used to obtain a partition of decision variables, given that a notion of similarity between the decision variables (or between groups of decision variables) is defined. In applications to transportation problems, the Euclidean or travel distance can be used as a similarity measure for clustering demand locations. For a given number of clusters K, the clustering algorithm iteratively finds a partition of demand locations into K clusters, in such a way each demand location belongs to the cluster whose center is the closest in distance. Another view of this approach is partitioning the network space into Voronoi cells, where, loosely speaking, the demand locations in each of the cell are close in distance. In Section 4.4.2, we discuss how the granularity of the partition, or the number of clusters, affects the performance of the sub-gradient algorithm.

Community Detection.

In this section we consider an approach that uses the structure of the optimization formulation itself to partition decision variables into blocks. We first build a network graph representing how decision variables (or groups of decision variables) are related through constraints. Then, we apply a community detection algorithm to find a partition informed by the structure of constraints of the optimization problem.

In the context of the transportation problem, we first build a network graph of demand locations. Consider a graph of n nodes, with each node representing one demand location.

Two nodes are connected by an edge if the corresponding demand locations are connected, that is, the two demand locations have access to a common supplier. The edge is weighted by the number of suppliers that can serve both of the locations, i.e., the number of constraints the two demand locations appear together. To this network, we apply a community detection algorithm to identify communities of demand locations where those in the same community are densely connected and those in different ones are sparsely connected. Then, we decompose the optimization problem according to the communities.

Among various algorithms developed in the community detection literature, we use the fast hierarchical agglomeration algorithm proposed by Clauset, Newman, and Moore [73]. The computational complexity of the algorithm is linear in the size of the network for many real-world networks. We briefly explain how the algorithm works below.

The community detection algorithm is based on a measure of a partition called the *modularity*, which evaluates how dense connections are within communities and how few there are between communities [74]. Before defining the measure, we introduce some notation. An n-by-n weighted adjacency matrix C is defined as

$$C_{vw} = \begin{cases} e_{vw} \text{ if nodes } v \text{ and } w \text{ are connected,} \\ 0 \text{ otherwise,} \end{cases}$$

where e_{vw} is the weight of the edge (v, w). Consider a partition of the nodes and for a node v, let c_v denote the community to which v belongs. Let $\delta(c_v, c_w)$ be 1 if $c_v = c_w$ and 0 otherwise. Let $m = \frac{1}{2}\sum_{v,w} C_{vw}$ be the sum of weights of all edges in the graph and let $k_v = \sum_w C_{vw}$ be the sum of weights of all edges from v. Then, the modularity of a partition is defined as:

$$Q = \frac{1}{2m} \sum_{v,w} \left(C_{vw} - \frac{k_v k_w}{2m} \right) \delta(c_v, c_w). \tag{4.8}$$

In the above definition, the fraction $k_v k_w/(2m)$ is the expected number of edges between

v and w where m edges are randomly assigned between nodes. Thus, the modularity measures how strong the community structure is over a random assignment of edges. More details of the modularity measure can be found in [73, 74, 68]. In practice, networks with the modularity greater than 0.3 appear to indicate significant community structure [68].

The community detection algorithm starts with a trivial division where each of the demand location forms a community. Then, it repeatedly joins two communities that results in the biggest increase of the modularity, until it reaches a partition where none of the joint operations improves the modularity score. More details of the algorithm can be found in [73].

4.3.2 Block Dual Decomposition

Each block in the output of the aforementioned variable partitioning algorithms may include multiple demand locations. Thus, the corresponding dual decomposition yields subproblems that include blocks of demand locations, and thus we call the proposed approach *block dual decomposition*. However, we emphasize that a grouping of decision variables vary by a partitioning method and that the community detection method groups those that are closely connected in the network graph representing the structure of the constraints. This characteristic of our approach makes the resulting sub-problems keep as much structure of the original optimization problem as possible, which is critical for the performance of the distributed sub-gradient method as explained in Section 4.2.3 and empirically shown in Section 4.4.

Let I_b for $b = 1, \ldots, B$ be the partition of demand locations given by the partitioning algorithms, thus satisfying $\cup_{b=1}^{B} I_b = I$ and $I_b \cap I_{b'} = \emptyset$ for $b \neq b'$. Let J_b be the set of supply locations that can serve the demand locations in I_b (e.g., within a pre-specified distance). Note that the set J_b's may not be disjoint as opposed to I_b's. For each block b, the suppliers that can serve only the demand locations in the block are said to be *interior suppliers* of block b, denoted as J_b^{in}, and the suppliers that are not interior suppliers but can

serve a demand location in I_b are called *boundary suppliers* of block b, denoted as J_b^{out}. Let $J^{\text{in}} = \cup_{b=1}^{B} J_b^{\text{in}}$ and $J^{\text{out}} = \cup_{b=1}^{B} J_b^{\text{out}}$. Note that J_b^{in} for $b = 1, \ldots, B$ are disjoint. For a demand location i, let $b(i)$ denote the block to which i belongs.

Consider the following Lagrangian relaxation of (GP):

$$\text{(BLR)} \min_{X_i} \; L(X, \Lambda) = \sum_{i \in I} \sum_{j \in J_i} w_{ij} x_{ij} + \sum_{j \in J^{out}} \lambda_j (\sum_{i \in I_j} x_{ij} - s_j) \tag{4.9}$$

$$\text{s.t.} \sum_{j \in J_i} x_{ij} \geq m_i \text{ for } i \in I, \tag{4.10}$$

$$\sum_{i \in I_j} x_{ij} \leq s_j \text{ for } j \in J^{in}, \tag{4.11}$$

$$X \geq 0. \tag{4.12}$$

Note that among the supply side constraints (4.3), only those corresponding to boundary suppliers were dualized in (BLR). Consequently, a fewer number of dual variables are needed than in the previous section. By following similar steps to those of the previous section, (BLR) is decomposed as follows:

$$\text{(BLR}_b\text{)} \min_{X_i} \sum_{i \in I_b} \sum_{j \in J_i} w_{ij} x_{ij} + \sum_{i \in I_b} \sum_{j \in J_b^{out} \cap J_i} \lambda_j x_{ij}$$

$$\text{s.t.} \sum_{j \in J_i} x_{ij} \geq m_i \text{ for } i \in I_b,$$

$$\sum_{i \in I_j} x_{ij} \leq s_j \text{ for } j \in J_b^{in},$$

$$X_i \geq 0 \text{ for } i \in I_b.$$

The resulting distributed subgradient algorithm is as follows.

Distributed Subgradient Algorithm with Block Dual Decomposition

1. Choose a starting point: $\Lambda^1 = \mathbf{0}$. Let $t := 1$.

2. Solve the local optimization problem (BLR_b) for each demand block b to obtain X_i^t for $i \in I_b$.

3. If (some stopping criterion) is satisfied, stop. Otherwise, $t := t + 1$, update the multipliers as below, and go to Step 2:

$$\lambda_j^{t+1} = \max\{\lambda_j^t + \alpha_t \left(\sum_{i \in I_j} x_{ij} - s_j \right), 0\} \text{ for } j \in J^{out}. \tag{4.13}$$

4.3.3 A General Approach

We have illustrated details of the block dual decomposition with community detection under the transportation problem setting. In this section, we present a similar approach, but with a broader applicability. Consider the following convex optimization problem:

$$\begin{aligned} &\min_{\mathbf{x}} f(\mathbf{x}) \\ &\text{s.t. } A\mathbf{x} \leq b, \end{aligned}$$

where $A \in R^{m \times n}$ and $f : R^n \to R$ is decomposable for each component of x, i.e., $f(\mathbf{x}) = \sum_{i=1,\ldots,n} f_i(x_i)$. For this general formulation, we illustrate how our approach can be applied to find a partition of decision variables for which the corresponding dual decomposition dualizes a minimal number of constraints.

Construct a graph in which each node represents a decision variable. Two nodes are connected if the two decision variables appear together in a constraint and the edge is weighted by the number of constraints they appear together. Note that in this section, each node represents a decision variable as opposed to the previous section where each node corresponds to a demand location for the transportation problem. Then, we apply the community detection algorithm to this network in order to identify a partition of decision

variables where connections within a group are dense but those between groups are sparse. A weighted adjacency matrix C is constructed as follows. We first form an indicator matrix $\widetilde{A} \in R^{m \times n}$ such that for all $i = 1, ..., m$ and $j = 1, ..., n$,

$$\widetilde{A}_{ij} = \begin{cases} 1 \text{ if } A_{ij} > 0 \text{ or } A_{ij} < 0 \\ 0 \text{ if } A_{ij} = 0. \end{cases}$$

Thus, $\widetilde{A}_{ij} = 1$ if x_i appears in constraint j. Then, an n-by-n weighted adjacency matrix C is defined as $C = \widetilde{A}\widetilde{A}^T$, thus C_{uv} is the number of times variables x_u and x_v appear in the same constraint, for all $u = 1, ..., n$ and $v = 1, ..., n$. Then, the modularity score of a partition is computed by using this C matrix as (4.8) and the community detection algorithm is applied. If the objective function is decomposable by groups of decision variables instead of each individual variable, then the aforementioned algorithm can be trivially extended by treating each block of variables as one node in the graph.

4.4 Numerical Results

In this section, we present experimental results for the distributed optimization framework with application to the transportation problem. We first explain the application and problem generation setup. Then, we empirically compare the sub-gradient method with the dual decomposition for each demand location (Section 4.2) and the three block approaches (Section 4.3) for problem instances with varying sizes and network structures. The distributed sub-gradient method was implemented in Julia, a high-performance programming language for numerical computing [75], along with Gurobi for optimization. For both the sequential and the parallel implementations, we used Intel Core Haswell Processors with 16 GB RAM on a Linux server with X86-64 bit architecture.

4.4.1 Problem Setup

In the experimental study, we generated problem instances based on an optimization model from a real application: matching children in need of primary care to healthcare providers in Georgia. The optimization problem takes the form of a general optimization model (GP) and it minimizes the total distance that patients have to travel to receive care.

In this problem setting, each demand location is a census tract and each supply location is a healthcare provider. Census tracts are used as proxies of communities and they form a contiguous division of a state. The patient population is aggregated at the census tract level using the geographic division established in the 2010 SF2 100% census data. In order to compute the number of children in each census tract, the 2012 American Community Survey data were used. Providers' practice location addresses, i.e., supply locations, were obtained from the 2013 National Plan and Provider Enumeration System (NPPES). More details about the application problem can be found in [76].

In the optimization problem, the decision variable x_{ij} denotes the number of children in demand location $i \in I$ assigned to supply location $j \in J$; w_{ij} is the distance between the demand location i and supply location j; m_i is the minimum number of patients needed to be served at demand location i; s_j is the maximum number of patients supply location j can accommodate. We allow the variables x_{ij} to be fractional for the computational tractability of the problem, and also because the number of children to be assigned from each location is typically large (approximately 2500-8000 children). In this application problem, there are 1955 demand locations and 3157 supply locations.

Using this optimization problem, we created problem instances with different sizes to evaluate the computational complexity of the proposed methodology with the problem size. First, we divided the map of Georgia into 50 blocks, 10 horizontally by 5 vertically, based on the longitudinal and latitudinal coordinates. Then, we counted the number of census tracts and provider locations in each block. For each block, we constructed a histogram of the demands (m_i's) of the census tracts in the block. We also obtained a histogram of the

supply capacities (s_j's) of providers in each block. Then, we generated a problem instance for a given number of demand and supply locations as follows. We determined the number of demand locations in each block in a way that the numbers of demand locations in different blocks of a generated instance are proportional to the numbers of demand locations in blocks of the original problem. In the same way we determined the number of supply locations in each block. Positions of demand and supply locations in each block were sampled randomly from the uniform distribution over the block. For each demand or supply location, the amount of demand or capacity was sampled from the empirical histogram of the block for demand or supply, respectively. In addition, a demand location i was said to have access to a supply location j if the distance w_{ij} between them is less than or equal to a given threshold $d_{\max}$. By changing the threshold $d_{\max}$ on the traveling distance, we were able to adjust the connectivity between demand and supply locations, thus changing the connectivity of the network. A lower $d_{\max}$ indicates a sparser network. A dummy supply location was included to guarantee feasibility.

4.4.2 Partitioning Methods

As an example of partitioning based on prior knowledge (Section 4.3.1), we grouped the census tracts based on their corresponding public health district affiliation. The Georgia Department of Public Health funds and collaborates with the public health districts throughout the state, while each health district oversees the operation of its affiliated health departments. There are 10 health districts in Georgia.

Similarly to partitioning based on prior knowledge, the clustering method (e.g. using k-means) uses the external (geographical) information for partitioning. Moreover, among the three partitioning methods considered in this paper, the clustering-based method is the only one for which a user can choose the number of blocks. The granularity of a partition is critical for the performance of the sub-gradient method, because it affects the number of iterations for the sub-gradient method to converge and the level of difficulty

of the sub-problems. In an extreme case where there is only one block, the sub-gradient method takes simply one iteration, but the sub-problem (which is the original problem) is the most complex comparing to that of any other partition. As a decomposition becomes finer, the resulting sub-problems become smaller, but the sub-gradient method requires more iteration to converge as we explained in Section 4.2.3.

Figure 4.2 illustrates this trade-off for the application considered. We generated a problem instance with 500 demand locations and 500 supply locations and the clusters were obtained using the k-means algorithm. The figure shows the CPU time to solve the largest sub-problem on average over different iterations and the number of iterations for the sub-gradient method to converge, for varying numbers of clusters. Note that in a synchronous distributed computing framework, the largest sub-problem is likely to be the bottleneck in each iteration. The time to solve the largest sub-problem was similar over different iterations. When k is greater than 9, the performance of the block dual decomposition becomes almost equivalent to the dual decomposition for individual demand locations, which converges in 2437 iterations.

This figure thus shows that the performance of the clustering-based partitioning method is affected heavily by the number of clusters and indicates that an optimal number of clusters must be determined specifically for a given problem instance. In the experiments in the next section, in order to fairly compare different partitioning methods, we choose k for the k-means to be the same as the number of blocks resulting from the community detection method.

The third variable partitioning approach, community detection, uses the optimization problem itself. We compare all three approaches in the next section.

4.4.3 Comparative Results

In this section, we generate multiple problem instances with varying size and complexity, and compare the empirical performance of the sub-gradient method with four decomposi-

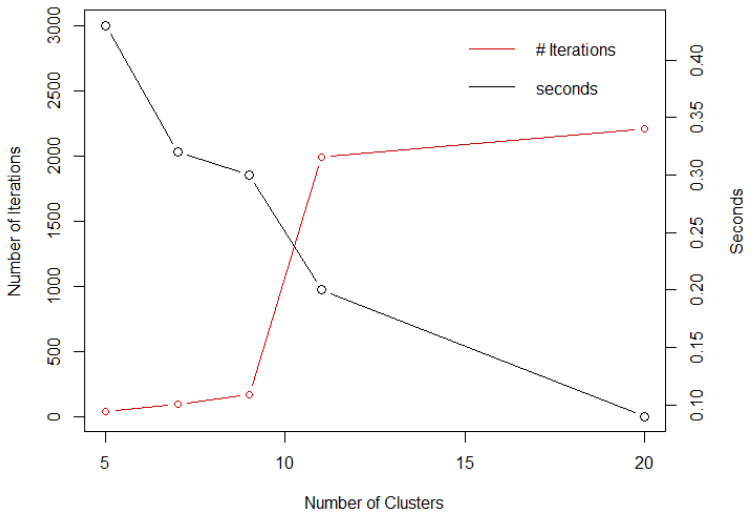


Figure 4.2: Trade-off between the number of iterations to reach convergence and the average time to compute the largest subproblem in each iteration.

tion approaches:

Baseline: Dual decomposition for each demand location;

Geo-HD: Block dual decomposition with the variable partitioning given by a geographic sub-regions (the health district grouping);

Geo-KM: Block dual decomposition with the variable partitioning given by k-means clustering of the demand locations using geographic distances; and

Opt-CD: Block dual decomposition with the variable partitioning given by the community detection based on the constraint structure of the optimization problem itself.

We first implemented the sub-gradient method with the variable partitioning methods in a sequential computing fashion, that is, all sub-problems in each iteration are solved sequentially using one computing node. In addition, we implemented a parallel version of the sub-gradient method for the block dual decomposition approaches in a distributed computing framework. The parallel implementation solves the sub-problems (BLR_b) in parallel at each iteration using three computing cores. The step size α_t was chosen to be c/t, where c is a constant scaling factor. For all of the methods, we considered different values of the scaling factor ($c = 1/10, 1/50, 1/25, 1/80$, and $1/100$), but all of the methods

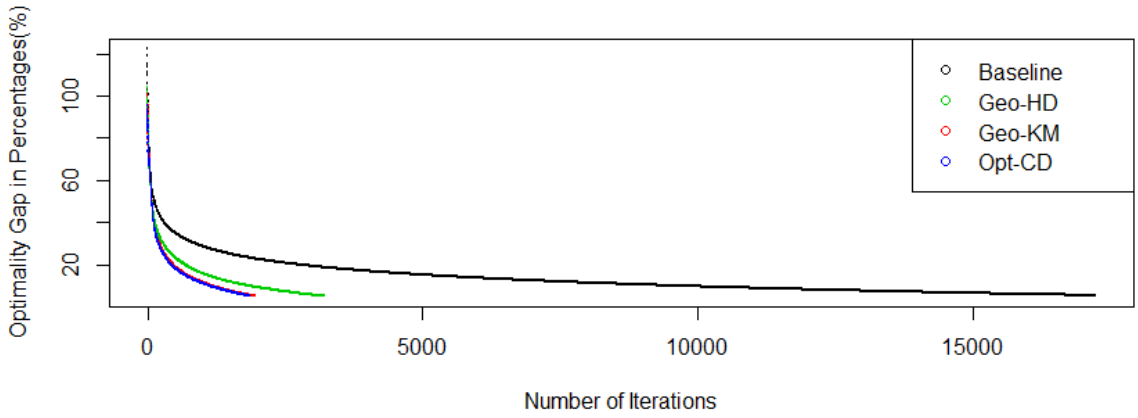


Figure 4.3: Comparison between the baseline dual decomposition and the block dual decomposition for 1000 demand locations and 1000 supply locations.

had the fastest convergence for the same c value at $1/25$, which we used for this comparison. We measured the number of iterations and the time in seconds (the CPU time for the sequential version and the wall clock time for the parallel version) required to reach a certain optimality gap percentage.

Figure 4.3 shows the comparison for a problem instance with 1000 demand locations, 1000 supply locations, and $d_{max} = 20$(miles) generated as explained in Section 4.4.1. The instance has 67,163 decision variables in total. We set $K = 6$ for the clustering-based approach to be consistent with the number of blocks yielded from the community detection algorithm. Figure 4.3 shows how the optimality gap progressed as a function of the number of iterations for sequential implementations of the baseline and the three block dual decompositions. Each method was terminated when the dual objective function value was within 5% of the true optimal objective function value.

Under any of the three partitioning methods, the block dual decomposition approach requires significantly fewer iterations to achieve the same optimality gap than the baseline decomposition. The *Baseline* method reached the stopping criterion after 17,231 iterations. On the other hand, *Geo-HD*, *Geo-KM*, and *Opt-CD* finished after 3234, 1965, and 1881 iterations, respectively. The *Baseline* took about 9.6 hours, but the *Geo-HD*, *Geo-KM*, and *Opt-CD* took 53, 26, and 37 minutes, respectively.

Table 4.1: Size of different problem instances.

		Instance 1	Instance 2	Instance 3
Problem Size	# Demand Locations	500	1,000	1,500
	# Supply Locaitons	500	1,000	1,500
	# Variables	26,444	67,163	228,000
	# Blocks with Geo-HD	10	10	10
	# Blocks with Geo-KM	7	6	6
	# Blocks with Opt-CD	7	6	6

Table 4.1 shows the size of additional problem instances, along with the number of blocks from the three partitioning methods. Table 4.2 shows similar comparisons on rate of convergence in number of iterations and computing time. Note that the community detection algorithm may result in a partition of the variables with many small communities or blocks. This is because many demand locations are in distant rural areas and have poor access to health care providers. For example, in the problem instance 2, the community detection algorithm yielded 10 blocks, with 4 small ones each of which has less than 20 demand locations. In Table 4.2, we only counted the communities with more than 20 demand locations. As the size of the problem grew, each algorithm took more iterations and more time to reach 5% optimality gap and the discrepancies between the methods also grew. The sequential version of the three block dual decomposition approaches achieved faster convergence in both the number of iterations and the run time than the *Baseline* by a large margin. The parallel implementations of the three block methods using three computing nodes yielded 2-2.5 times speed up comparing to their sequential counterparts.

In order to evaluate how the connectivity of the network graph affects the performance of the partitioning methods, we constructed problem instances with 1500 demand locations and 500 supply locations, but different values of $d_{\max} = 20$, 25, and 30 in miles. Recall that $d_{\max}$ is the maximum distance to travel and that a supply location is accessible from a demand location if the distance between them does not exceed $d_{\max}$. Thus, the larger $d_{\max}$ is, each supply constraint involves more decision variables. In the network of demand locations for community detection (defined in Section 4.3.1), two demand locations

Table 4.2: Comparison on reaching 5% optimality gap for problem instances with varying sizes.

		Instance 1	Instance 2	Instance 3
# Iterations	Baseline	4,375	17,231	>8,000
	Block with Geo-HD	568	3,233	3,164
	Block with Geo-KM	163	1,965	4,122
	Block with Opt-CD	345	1,881	1,555
Run Time (in Seconds)	Baseline	1,723	34,532	>300,000
	Block with Geo-HD	212	7,546	12,959
	Block with Geo-KM	72	3,846	17,988
	Block with Opt-CD	133	5,115	6,016
	Dist Block Geo-HD	83	3,210	6,712
	Dist Block Geo-KM	41	1,577	9,144
	Dist Block Opt-CD	52	2,243	2,956

are connected if they share an accessible supply location, and thus, a larger value of $d_{\max}$ implies a more dense network. Figure 4.4 compares the baseline and the three approaches using block partitioning up to 2500 iterations for the three instances with different connectivity. For the three values of $d_{\max} = 20$, 25, and 30, each demand location had access to 51, 68 and 84 providers on average, respectively.

We observe that the approaches using block partioning perform better than the baseline consistently for different values of $d_{\max}$, but the discrepancy of performance gets larger as the threshold increases. For $d_{\max} = 30$ at the 2500th iteration, the optimality gap is 39% for the baseline, and 32%, 19%, and 15% for the block with *Geo-HD*, *Geo-KM*, and *Opt-CD*, respectively. For $d_{\max} = 25$, the optimality gap is 31% for the baseline, and 27%, 9%, and 10% for the block with the three partition methods respectively. For $d_{\max} = 20$, the optimality gap is 19% for the baseline, and 11%, 2%, and 6% for the block with the three partition methods respectively. Thus, the performance of the block approaches, and also that of the community detection approach, is significantly improved as the network becomes more dense.

The effect of the network structure on the performance can be explained geometrically as follows. For a larger value of the threshold, each provider is accessible from more demand locations and thus, each provider constraint contains more decision variables. In

that case, dualizing each provider constraint results in a bigger change on the feasible region in the following sense. For example, consider the following two relaxations: relaxing $x_1 + x_2 \leq 1$ from $\{(x_1, x_2) \mid x_1 + x_2 \leq 1, x_1 \geq 0, x_2 \geq 0\}$ and relaxing $x_1 \leq 1$ from $\{(x_1, x_2) \mid x_1 \leq 1, x_1 \geq 0, x_2 \geq 0\}$. The former can be viewed to yield a bigger change than the latter. In this sense, when $d_{\max}$ is larger, dualizing each provider constraint causes a bigger change on the feasible region. Moreover, note that the baseline dual decomposition dualizes more provider constraints than the proposed approach. Therefore, as $d_{\max}$ increases, the discrepancy between the feasible regions of the Lagrangian relaxation and of the original problem becomes more significant for the baseline than it does for the block approaches. Thus, when $d_{\max}$ increases (i.e., the network gets more dense), the baseline performs more poorly as compared to the block approaches.

The block dual decomposition with different partitioning methods consistently outperform the baseline dual decomposition in all of our experiments. It is however worthwhile to note that, although k-means seems to produce better results among the three in most settings, it highly depends on the inherent problem structure, and should be compared on a case by case basis.

4.5 Conclusion

In this paper we proposed a novel approach for determining a partition of decision variables while decomposing a large-scale optimization problem in a way that improves the performance of distributed optimization methods. We first showed that the partitioning of the decision variables in dual decomposition could be crucial for the empirical performance of a distributed sub-gradient method. Then, we proposed three methods for finding a partition of variables that minimizes the number of constraints being dualized. Our method groups variables that should be in the same sub-problem in order to achieve the best performance of distributed methods.

The experimental study using the real application shows that the proposed approach

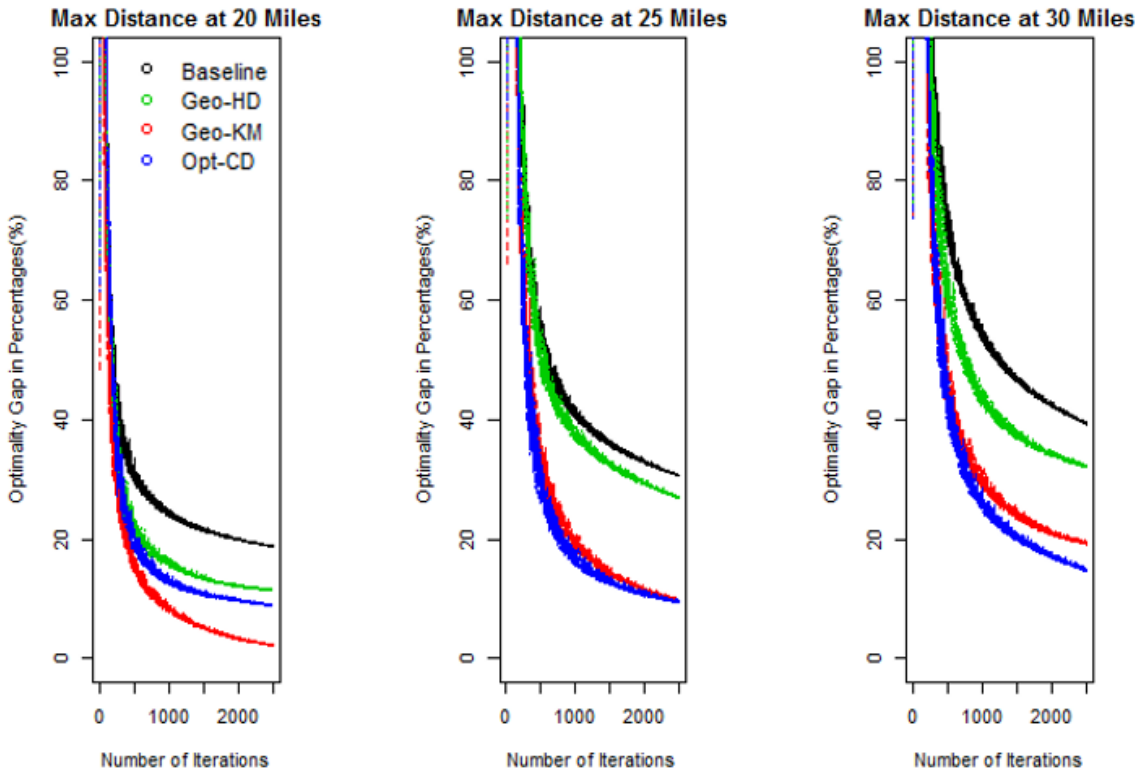


Figure 4.4: Comparison on rate of convergence between the dual decomposition and the block dual decomposition with varying network structures.

can be used to find a partition for dual decomposition that speeds up the convergence of distributed sub-gradient methods and that the performance of our approach is significantly better as each constraint involves more variables and thus, the connectivity among the variables gets stronger. In addition, the proposed methodology can be easily combined with other established techniques that improve the rate of convergence, such as incremental methods [77], smoothing techniques [78, 38], adaptive subgradient methods [79] among others.

We highlight here that research in computer science has introduced approaches and algorithms for sub-problem decompositions in a way that communication between computing nodes are minimized [80, 81, 82, 83]; however, there are some key differences between those works and this paper. First, our goal for finding a decomposition is not to minimize communication but to minimize the number of constraints being dualized, in order to conserve as much structure of the original problem as possible while decomposing the optimization problem. Also, each node in the network of this paper represents not a computing node but a decision variable or a group of decision variables (such as a demand location in the transportation problem). While we focused on speeding up the convergence of the distributed method in this paper, the proposed methodology may also be used for minimizing communication between computing nodes, which is a future research topic.

Another potential future research is examining whether the proposed variable partitioning method can be applied to other distributed optimization approaches. Coordinate descent methods [54, 56, 84] have gained popularity recently. In coordinate descent methods, variables are partitioned into groups, one of which is chosen to be updated in each iteration. Thus, the approach proposed in this paper can also be used to find a partition for coordinate descent methods; which may also benefit from the community structure of decision variables found by our approach.

One limitation of the proposed approach is load balancing. None of the introduced partitioning methods guarantees sub-problems with equal sizes. A very large block that

dominates the execution time can effect the level of speedup due to synchronization. One approach is to design a partition algorithm that penalizes on the load imbalance. Another direction is to pursue an asynchronous version of the sub gradient algorithm to reduce the impact of having heterogeneous sub problem size.

CHAPTER 5

CLUSTERING THE PREVALENCE OF PEDIATRIC CHRONIC CONDITIONS IN THE UNITED STATES USING DISTRIBUTED COMPUTING

In this section, we presents an approach to clustering the prevalence of chronic conditions among children with public insurance in the United States. The data consist of prevalence estimates at the community level for 25 pediatric chronic conditions. We employ a spatial clustering algorithm to identify clusters of communities with similar chronic condition prevalences. The primary challenge is the computational effort needed to estimate the spatial clustering for all communities in the U.S. To address this challenge, we develop a distributed computing approach to spatial clustering. Overall, we found that the burden of chronic conditions in rural communities tends to be similar but with wide differences in urban communities. This finding suggests similar interventions for managing chronic conditions in rural communities but targeted interventions in urban areas.

5.1 Introduction

The Medicaid public insurance program covers more than 36 million children in the United States yearly [85]. Children covered under this program are from low-income families or/and with severe health disabilities. Disparities in health outcomes for Medicaid-enrolled children are substantive and of great concern nationally [86]. A first step in addressing such disparities is measurement and evaluation of the health outcomes for this population. Towards this objective, in this paper, we study the burden of chronic conditions in the Medicaid-enrolled child population, which can vary across communities within each state and across states. Characterizing the burden of chronic conditions can help in identifying communities with most need for interventions for improving health outcomes.

We compiled a unique (in-treatment) prevalence data on multiple chronic conditions

common in children. The prevalence data are derived from patient-identifiable medical claims from the 2011 Medicaid Analytic eXtract (MAX) files acquired from the Centers for Medicare and Medicaid Services (CMS). The prevalence data are census tract estimates of the percentage of Medicaid-enrolled children diagnosed with a chronic condition, with a total of 64,873 census tracts across the United States, and 25 chronic conditions. The objective in this study is to characterize the burden of chronic conditions in communities by using a clustering or segmentation of the population of children based on the level of prevalence of their chronic conditions. This clustering approach reduces the information content in such large prevalence data into simple data clustering summaries by borrowing information across all census tracts (proxies of communities) and across prevalent childhood chronic conditions. The end point is to create a clustering map of the burden of chronic conditions, which can be used in informed decision making and targeted healthcare interventions.

An important challenge in deriving a clustering for the prevalence data is the presence of strong spatial dependence. Spatial dependence arises because proximal communities will have similar levels of chronic conditions; proximal communities will have similar demographics, social-economics and environmental factors, which can influence the development and the severity of chronic conditions [87, 88]. These types of spatial effects have been widely modeled in disease mapping. Reviews of methodology for spatial epidemiological data in general may be found in [89, 90, 91, 92]. Most models were developed in spatial smoothing and regression settings. Incorporating spatial dependence is vital in understanding geographical patterns in disease incidence and mapping.

One first attempt for detecting spatial point clusters and hotspots using exploratory methods is the Geographical Analysis Machine (GAM) developed by [93] and later improved by [94]. Another popular method is to encode the spatial dependence (or other form of dependence) into the feature space. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [95], Ordering Points To Identify the Clustering Struc-

ture(OPTICS) [96], and other related variations [97, 98] are the most well-known density based clustering algorithms for correlated data, where distance based measures are used to allocate data to clusters. [99] transformed raw pixel data in images into a joint color-texture-position feature space, and used the Expectation Maximization for Gaussian Mixtures to construct a small set of image regions that are coherent in color and texture. [100] introduced a model-based method for clustering random time-varying functions under spatially interdependence. [101] extended the hidden Markov models to the spatial domain with a finite-mixture model for Poisson rates, where the mixture component follows a spatially correlated process, the Potts model. This model is flexible in terms of assumptions but may be cumbersome to implement on extremely large data.

In this paper, we apply a general spatial clustering approach to cluster high-dimensional data assuming spatial dependence in the observed response data, while not restricting the spatial effect to be the same across the spatial domain. The main challenge in implementing this spatial clustering method to the nationwide prevalence data is the computational effort; the method is not scalable to a large number of spatially-dependent responses. To address this challenge, in this paper, we develop a distributed-computing spatial clustering method.

Distributed computing has become a much needed alternative modeling approach in many research domains, particularly in statistical learning, due to the advent of large size and complex datasets. The size of the data collected are sometimes too large to be stored at a central location, and the level of computation needed for statistical learning may not scale up to the data dimensionality. In addition, data in some cases are naturally collected in a decentralized fashion at a local level, and communication between local servers and a central machine is expensive and wasteful. The data are usually assumed to be independent to alleviate the computational burden, since data in each node can be calculated separately in a distributed fashion. [102] indicated that algorithms applied to independent data are easily parallelizable on multicore computers, in a Map Reduce framework. [103] provided a general framework for distributing expectation-maximization algorithms under indepen-

dence of the response data in which not only is the computation distributed, but the storage of parameters and expected sufficient statistics is also fully distributed. However, when strong are present among the response variables, the independence assumption is therefore violated, which can produce misleading inferences.

One contribution of this paper is thus the derivation and development of a distributed computing solution to the estimation of the clustering model under spatially interdependence. The estimation approach requires innovation in the decomposition of the log-likelihood function in a way that its maximization can be distributed across multiple computing cores. A second contribution in this paper is that we not only derive the distributed estimation approach but also implement it within the applied problem, specifically, identifying geographic clusters of the burden of pediatric chronic conditions, where each cluster can be characterized by different prevalence levels of chronic conditions and by different groups of the conditions.

In the following section, we present the approach for deriving the prevalence data. In the section that follows, we will continue with the introduction of the general form of the expectation-maximization (EM) algorithm in solving Gaussian Mixture Models, then we relax the independence assumption by re-formulating the E-step and M-step, and proposing an efficient parallel EM Algorithm. We apply the proposed algorithm to deriving the clustering map of the chronic disease prevalence among children enrolled in Medicaid using the large-scale prevalence data. We conclude with a discussion on the implications of the clustering map towards targeted healthcare interventions.

5.2 Chronic Condition Prevalence for the Medicaid-enrolled Children

5.2.1 Data Source

We analyze the patient-level claims from the 2011 Medicaid Analytic eXtract (MAX) files obtained from the Centers for Medicare and Medicaid Services (CMS). The research in this study was approved by CMS (Data Use Agreement #23621) and by the Institutional

Review Board of Georgia Tech (protocol #H11287). All data derived from the MAX files meet a minimum cell size of 11 in terms of number of patients according to the Data Use Agreement with CMS. We focus on children age 0 to 17.

5.2.2 Prevalence Estimation

We derive the prevalence estimates using the 3M Clinical Risk Grouping software [104]. Episode Diagnostic Categories (EDCs) are derived for each child enrolled in Medicaid using the child's diagnosis codes, procedure codes, and national drug codes (NDCs) found in the recorded medical claims in the MAX files. EDCs are used to determine a patient's Primary Chronic Disease, which is the most significant chronic disease actively being treated, and its severity for each organ system.

We consider EDCs for the following 25 conditions: Acute Bronchitis and Bronchiolitis, Acute Respiratory Diagnoses - Moderate, Acute Skin Diagnoses, Acute Stress and Anxiety, Attention Deficit Hyperactivity Disorder (ADHD), Allergies, Asthma, Autism, Bipolar, Chronic Mental Health, Chronic Stress, Conduct and Behavior, Dental Diagnoses, Depression, Depressive and Other Psychoses, Developmental Language Disorder, Developmental Speech and Learning, Diabetes2, Epilepsy and Epilepsy Complex, Major Mental Health, Psoriasis, Schizophrenia, Social Problems, Upper Respiratory Infections. These conditions were selected due to their high prevalence among children enrolled in the Medicaid program. According to the data use agreement with CMS, we cannot disclose any information when the cohort population is less than 11 patients, thus lower prevalence conditions cannot be captured in our analysis.

For each condition or EDC, we obtained the population of Medicaid-enrolled children with the condition along with the number of enrollment months of these children within each zip code and county. We derived the prevalences of conditions by dividing the total number of member months of patients treated for a given condition by the total number of member months of all children on Medicaid for each county and zip code area. We further

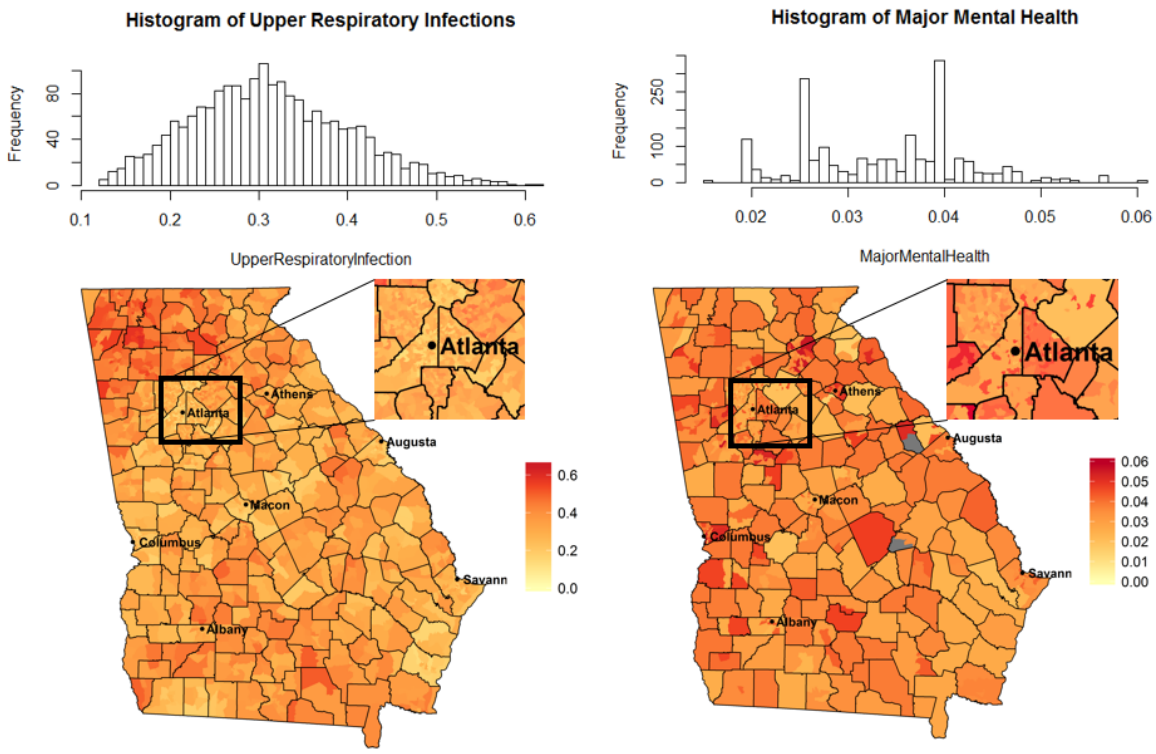


Figure 5.1: Histogram and heat map of prevalence for upper respiratory infections and major mental health in the state of Georgia.

estimated the census tract prevalence using the zip code and county estimates along with geographic information of the boundaries of the different geographic devisions (county, zip code, census tracts) and the information on the population count across the geographic divisions. For cells with less than 11 patients, we used the mean estimation at the state level, along with a generated beta noise term.

Overall, we have a total of 64,873 census tracts for which we have obtained prevalence estimates for the 25 conditions. The census tracts cover the entire United States excluding Colorado and Idaho due data unavailability. The prevalence of the EDCs and their denomination were provided by the 3M Clinical Risk Grouping software and the derivation of the census tract prevalence estimates were based on the MAX claims data.

5.2.3 Exploratory Analysis

The prevalence across the 25 chronic conditions varies widely, with Epilepsy as the least prevalent condition (ranging from 0.2% to 0.8%) and upper respiratory infections as the most prevalent condition (ranging from 12.6% to 61.3%). Figure 5.1 shows the histogram and heat map of the prevalence for upper respiratory infections and major mental health in the state of Georgia. The distribution for the upper respiratory infections is approximately uni-modal but for the major mental health condition it is multi-modal. The heat map shows the presence of spatial dependencies, where nearby geographical locations tend to have similar level of prevalence. The strength of spatial dependence however differs from region to region and by condition. In urban locations, such as the Atlanta metropolitan area, the census tracts tend to be much smaller and denser. The prevalences are more similar in these areas than in rural locations where census tracts are larger and further away. For more prevalent conditions, such as the upper respiratory infections, the prevalence across the map is smoother than for less prevalent conditions, thus differences between nearby census tracts are relatively small.

5.3 Statistical Modeling Using Distributed Computing

5.3.1 Nominal EM Algorithm for Gaussian Mixture Models

The Expectation Maximization (EM) Algorithm is a class of iterative methods for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobservable or latent variables [105]. Each EM iteration alternates between performing an expectation (E) step, which updates the expectation of the log-likelihood function evaluated using the current estimates for the parameters, and a maximization (M) step, which estimates parameters maximizing the expected log-likelihood given the input from the E step of the previous iteration. The EM is frequently used for modeling mixtures of distributions, where data are commonly assumed to be generated from mixtures

of multivariate Gaussian distributions (GMM) assuming unknown number of mixtures and unknown mixture weights. Modeling the mixture of distributions can also be viewed as a data clustering method [106, 107].

The observed response data are $Y_1, Y_2, ..., Y_N$ where Y_i is a p-dimensional vector of measurements, in this paper, the prevalence estimates for the 25 pediatric chronic conditions. N is the number of responses. We further assume that the distribution of Y_i is a realization from a finite mixture model with C components:

$$p(y|\Theta) = \sum_{k=1}^{C} w_k p_k(y|\theta_k)$$

- $p_k(y|\theta_k)$ is the k-th mixture component where this mixture is identified by the parameter θ_k. For mixtures of Gaussians, $\theta_k = \{\mu_k, \Sigma_k\}$ are the mean and the covariance specifying the k-th Gaussian.

- w_k are the mixture weights, representing the probability that a randomly selected Y was generated by component k.

The unobserved data are the latent variables $Z_1, Z_2, ..., Z_N$ where Z_i has a C-dimensional multinomial distribution specifying the cluster membership of Y_i. Thus given $Z_{ik} = 1$ and $Z_{ic} = 0$ for $c \neq k$ where k takes values in $\{1, 2, \ldots, C\}$, Y_i has a distribution with the density function $p_k(y|\theta_k)$.

The EM algorithm is an iterative algorithm that starts from some initial estimates of $\Theta = (\theta_1, \ldots, \theta_C)$ and of $\mathbf{w} = (w_1, \ldots, w_C)$, and then proceeds to iteratively update Θ and $\mathbf{w}$ until convergence. Each iteration consists of an E-step at which we update the mixture weights and impute the cluster memberships and an M-step at which we estimate Θ given the imputed cluster membership.

For classic mixtures of multivariate distributions, the responses to be clustered are generally assumed independent and hence the EM algorithm can be distributed easily across multiple computing nodes [103]. However, in this paper, we assume the response data are

spatially interdependent.

5.3.2 Correlation Structure

The proposed spatial EM algorithm extends the nominal EM algorithm (under independence assumption) by incorporating spatially correlated random errors. In our application, the spatial correlation structure is a function of the proximity between pairs of census tract centroids, assumed to be defined by a Matérn correlation function, which is widely used in spatial statistics and geostatistics [108, 109, 110].

The most granular information we have for each patient is the residential zip code, not the exact address. Therefore, we are treating the prevalence estimates as point masses at the centroids of each area unit, instead of a point process across the geographic area. Alternatively, we can use a power law on the order or proximity of the neighborhoods, such that a small neighboring region would be attributed a stronger link than a large neighbor with its centroid further apart.[111] However, as census tracts are defined based on settlement density, it is desirable that dense areas, mostly consisting of small neighboring regions, have stronger spatial dependencies than rural areas, where large neighbors' centroids are further apart.

Furthermore, instead of considering spatial correlation between every possible pair of locations, which deems to be intractable, we enforce a neighborhood structure - that is, for each location, we only consider the closest $M-1$ number of neighbors, resulting in a neighborhood of size M. M can be fixed, or can vary for different response i. For example, we can assign a larger M to urban locations than to rural locations, since the spatial effect is expected to be stronger. An alternative is to sparsify the spatial correlation matrix by setting a hard threshold. In addition, the correlation is assumed to decay exponentially as the distance between two locations increases. Other correlation structures can be considered but for simplicity of the interpretation and implementation, we use classic approaches to specify the correlation structure.

The neighborhood criterion is similar to spatial tampering and the Gaussian Markov random fields (GMRF) advocated in [112, 113, 114]. Under GMRF modeling, observations at each location only depend on a set of neighbors. [113, 114] GMRF can efficiently model most of the spatial covariance functions. [115] The method has been extended to integrate a stochastic partial differential equation approach and integrated nested Laplace approximations (INLA) for faster computation, which has been implemented in R. [116, 114, 117] To model irregular grids, the space is divided into a collection of the so-called Delaunay triangulations, and is approximated using basis functions. While this method applies effectively to moderate size dataset, an application with a large number of spatial points, e.g. 64,873 locations as for the U.S. prevalence data, can easily require more than 4 million basis functions to estimate.

We also assume the features are uncorrelated. This can be achieved by assuming independence on the feature space; to achieve independence between the features, the feature set can be preprocessed into uncorrelated orthogonal basis set, for example, using the Principal Component Analysis (PCA), which is common practice [118]. In the next section, we will see that the assumption of independence on the feature space significantly reduces the computational complexity and makes the algorithm parallelizable.

5.3.3 Expectation Step

In the E-step, we evaluate the expected cluster membership probability for each response based on the parameters estimated in the M-step. In the derivations below, since the parameters specifying the mixtures $\Theta = (\theta_1, \ldots, \theta_C)$ are assumed fixed in the E-step (provided by the estimates derived in the M-step), we drop the conditioning on the set of parameters Θ for ease of illustration.

Conditional Model: The model for the i-th response or measurement is:

$$Y_i|(Z_{ik} = 1) = \mu + \mu_k + s_i + e_i$$

where Z_i is the latent variable (cluster membership) for the response i, μ is the global mean, μ_k is the cluster mean for cluster k, with $\sum_{k=1}^{k=C} \mu_k = 0$, where C is the total number of clusters, the spatial random effect s_i, and the independent error term e_i. $\mathbf{Y}$ denotes the vector of all responses and Y_i denotes the i^{th} response; the k^{th} membership probability for the response i is denoted as w_{ik}.

Under interdependence among the responses, the estimation of w_{ik} involves more complex computations:

$$\begin{aligned} w_{ik} &= E[Z_{ik} = 1|\mathbf{Y}] = P(Z_{ik} = 1|\mathbf{Y}) \approx P(Z_{ik} = 1|Y_i, Y_{N(i)}) \\ &= \frac{P(Z_{ik} = 1)f(Y_i, Y_{N(i)}|Z_{ik} = 1)}{\sum_{c=1}^{c=C} P(Z_{ic} = 1)f(Y_i, Y_{N(i)}|Z_{ic} = 1)} \end{aligned}$$

where $N(i)$ denotes the set of indexes of the responses that are neighbors of the i^{th} response. In this case, the dependence structure among the responses is encoded in the parameter estimations of the latent classes, which we will discuss in more detail in the M step. Therefore, the expected membership probability for sample i depends on its neighbors, i.e., the probability density function $f(Y_i, Y_{N(i)}|Z_{ik} = 1)$ needs to be calculated jointly. In what follows, we will focus on how to estimate this joint probability efficiently. For ease of presentation, we will use μ to represent $\mu + \mu_k$.

Denote the m^{th} neighbor of response i as $Y_{N(i,m)}$, where $m = 1, 2, ...M-1$. Calculating the joint probability of this neighborhood can be computationally intense and not scalable, since only $Z_{ik} = 1$ is given, but not its neighbor's cluster memberships. The joint density is calculated as:

$$\begin{aligned} & f(Y_i, Y_{N(i)}|Z_{ik} = 1) = f(Y_i, Y_{N(i,1)}, ..., Y_{N(i,M-1)}|Z_{ik} = 1) \\ = & \sum_{k_{N(i,1)}=1}^{k_{N(i,1)}=C} f(Y_i, Y_{N(i,1)}, ..., Y_{N(i,M-1)}|Z_{ik} = 1, Z_{N(i,1)k_{N(i,1)}} = 1) \times \\ & P(Z_{N(i,1)k_{N(i,1)}} = 1) \end{aligned}$$

We denote the mixture weight $P(Z_{N(i,m)k_{N(i,m)}} = 1)$ with $w_{N(i,m)k_{N(i,m)}}$. We then expand the joint density function for all responses in the neighborhood:

$$f(Y_i, Y_{N(i)}|Z_{ik} = 1) = \sum_{k_{N(i,1)}=1}^{k_{N(i,1)=C}} w_{N(i,1)k_{N(i,1)}} ... \sum_{k_{N(i,M-1)}=1}^{k_{N(i,M-1)=C}} w_{N(i,M-1)k_{N(i,M-1)}} \times$$

$$f(Y_i, Y_{N(i,1)}, ..., Y_{N(i,M-1)}|Z_{ik} = 1, Z_{N(i,1)k_{N(i,1)}} = 1, ..., Z_{N(i,M-1)k_{N(i,M-1)}} = 1)$$

In each summation, the joint density function are conditioned on the membership of the response i and its neighbors. However, the amount of computation doesn't scale with the increasing size of the neighborhood, as the joint density function needs to be expanded in the neighborhood of each response and in each cluster, which results in C^M joint density estimations. One alternative is to perform a hard clustering on $w_{N(i,m)k_{N(i,m)}}$ $\forall m = 1, 2, ..., M-1$ such that $w_{N(i,m)k^*_{N(i,m)}} = 1$ for $k^*_{N(i,m)}$ which maximizes over all $k_{N(i,m)}$ and $w_{N(i,m)k_{N(i,m)}} = 0$ for all other $k_{N(i,m)}$. Thus we have the following approximation

$$f(Y_i, Y_{N(i)}|Z_{ik} = 1) \approx f(Y_i, Y_{N(i,1)}..., Y_{N(i,M-1)}|$$

$$Z_{ik} = 1, Z_{N(i,1)k^*_{N(i,1)}} = 1, ..., Z_{N(i,M-1)k^*_{N(i,M-1)}} = 1)$$

An interpretation of the approximation above is as follows: the memberships of the neighboring responses are assumed to be fixed based on the membership matrix calculated in the previous M step, and only the membership of response i varies. This heuristic is similar to successive methods such as backfitting and Gauss-Seidel. We denote $f(Y_i, Y_{N(i,1)}..., Y_{N(i,M-1)}|Z_{ik} = 1, Z_{N(i,1)k^*_{N(i,1)}} = 1, ..., Z_{N(i,M-1)k^*_{N(i,M-1)}} = 1)$ as $f(\mathbb{Y}_i|Z_{ik} = 1)$, where $\mathbb{Y}_i$ is a M-by-p matrix. Denote the M-by-M matrix S_i as the spatial covariance matrix for the neighborhood around i, and the p-by-p matrix Σ_i as the covariance matrix for the random error ϵ_i for response i, i=1,2,...,M, where Σ_i is a diagonal matrix with the diagonal provided by [σ^2_{i1},...,σ^2_{ip}]. The neighborhood $\mathbb{Y}_i$ thus follows a matrix normal distribution, whose variance is the Kronecker product of the S_i and the corresponding Σ_i for each response in the

neighborhood.

We further decompose:

$$\mathbb{Y}_i|(Z_{ik}=1) = \mu + S_i^{\frac{1}{2}}A$$

where each row l of A is independent, $A_{l\cdot} \sim N(\mu_l, \Sigma_l)$. We then have:

$$f(\mathbb{Y}_i|Z_{ik}=1) = \prod_{l=1}^{l=M} \frac{1}{\sqrt{det(S_i)}} f(A_{l\cdot}|\mu=\mathbf{0}, \Sigma=\Sigma_l)$$

$f(\mathbb{Y}_i|Z_{ik}=1), \forall i=1,...,n, k=1,...,C$ can be computed using distributed computing, for each i separately or for groups of i's. This can further be used in estimating the cluster weights w_{ik}, concluding the E-Step.

5.3.4 Maximization Step

In the Spatial EM algorithm, the parameter set Θ contains of $(\mu_k, \Sigma_k) = \theta_k$, for all $k = 1,...,C$. It is however computationally challenging to obtain the MLEs for these parameters when there is dependence in the sample data. Alternatively, we can use the Maximum Pseudo-likelihood Estimation [119, 120]

$$maxE[l(\Theta;Y)] \quad = \quad max \sum_{i=1}^{i=n} \sum_{k=1}^{k=C} w_{ik} log f(Y_i|Y_{N(i)}, Z_{ik}=1)$$

We use a similar technique as used in the computation in the E-step to account for the dependence structure. For each response i belonging to cluster k, we have:

$$X_i = (Y_i - \mu_k)\Sigma_k^{-\frac{1}{2}}$$

$$X_{N(i)} = \sum_{l=1}^{l=M-1} e_l \sum_{k=1}^{k=C} w_{ik}(Y_{i_l} - \widetilde{\mu}_k)\widetilde{\Sigma}_k^{-\frac{1}{2}}$$

where e_l is a vector of length M-1, where the l^{th} element is 1 with all other values being zero, $\widetilde{\mu}_k$ and $\widetilde{\Sigma}_k$ are parameters estimated in the previous iteration of the EM algorithm. We then have the following:

$$max \quad \sum_{k=1}^{k=C} \sum_{i=1}^{i=n} w_{ik} log f(Y_i | Y_{N(i)}, Z_{ik} = 1) =$$
$$max \quad \sum_{k=1}^{k=C} \sum_{i=1}^{i=n} \sum_{j=1}^{j=p} w_{ik} log(\frac{1}{\sigma_{kj}} f(X_{ij} | [X_{N(i)}]_{\cdot j}, Z_{ik} = 1))$$

Expand the spatial correlation matrix S_i as:

$$\begin{bmatrix} S_{i_{11}}, S_{i_{12}} \\ S_{i_{21}}, S_{i_{22}} \end{bmatrix}$$

where $S_{i_{11}}$ is 1, the vector $S_{i_{12}}$ of length $M-1$ is the correlation between response i and its neighbors, $S_{i_{21}}$ is the correlation between response i's neighbors and itself, and the (M-1)-by-(M-1) matrix $S_{i_{22}}$ is the correlation matrix among response i's neighbors. We then have $X_{ij} | [X_{N(i)}]_{\cdot j}, Z_{ik} = 1 \sim N(\bar{\mu}_{ij}, \bar{\Sigma}_{ij})$, where

$$\bar{\mu}_{ij} = S_{i_{12}} S_{i_{22}}^{-1} [X_{N(i)}]_{\cdot j}$$

$$\bar{\Sigma}_{ij} = 1 - S_{i_{12}} S_{i_{22}}^{-1} S_{i_{21}}$$

Therefore, we have:

$$max \quad \sum_{k=1}^{k=C} \sum_{i=1}^{i=n} \sum_{j=1}^{j=p} w_{ik} log(\frac{1}{\sigma_{kj}} f([X_{ij} | [X_{N(i)}]_{\cdot j}, Z_{ik} = 1)) =$$
$$max \quad \sum_{k=1}^{k=C} \sum_{i=1}^{i=n} \sum_{j=1}^{j=p} w_{ik} \{-log(\sigma_{kj}) - \frac{(\frac{Y_{ij} - \mu_{kj}}{\sigma_{kj}} - \bar{\mu}_{ij})^2}{2\bar{\Sigma}_{ij}}\} = G(Y)$$

Setting the first derivatives of the pseudo-likelihood to zero, we get the following estimation:

$$\widehat{\mu}_{kj}^{mple} = \frac{\sum_{i=1}^{i=n} \frac{w_{ik}}{\bar{\Sigma}_{ij}} (Y_{ij} - \bar{\mu}_{ij}\sigma_{kj})}{\sum_{i=1}^{i=n} \frac{w_{ik}}{\bar{\Sigma}_{ij}}},$$

and σ_{kj}^{mple} is the positive root of the following quadratic equation:

$$\sum_{i=1}^{i=n} w_{ik}\sigma_{kj}^2 - \sum_{i=1}^{i=n} \frac{w_{ik}}{\bar{\Sigma}_{ij}} (Y_{ij} - \mu_{kj})\bar{\mu}_{ij}\sigma_{kj} + \sum_{i=1}^{i=n} \frac{w_{ik}}{\bar{\Sigma}_{ij}} (Y_{ij} - \mu_{kj})^2 = 0$$

We initialize $\widehat{\mu}_{kj}^{mple}$ and $\widehat{\sigma}_{kj}^{mple}$ with the estimates from the previous iteration, and solve the equations iteratively.

If the correlation among samples is minimal, the spatial correlation matrix $S_i, \forall i = 1, ..., n$ becomes diagonal, with $\bar{\mu}_{ij} = 0$ and $\bar{\Sigma}_{ij} = 1$; therefore, we have:

$$\begin{aligned} \mu_{kj} &= \frac{\sum_{i=1}^{i=n} w_{ik} Y_{ij}}{\sum_{i=1}^{i=n} w_{ik}} \\ \sigma_{kj}^2 &= \frac{\sum_{i=1}^{i=n} w_{ik} (Y_{ij} - \mu_{kj})^2}{\sum_{i=1}^{i=n} w_{ik}}, \end{aligned}$$

which coincides with the estimation based on the nominal EM algorithm with independent responses. The proposed method is therefore a generalization of the nominal EM algorithm.

5.3.5 Model Selection

Similar to most of the model-based clustering algorithms, the number of clusters needs to be finely tuned to obtain a set of meaningful clusters. Common variable selection methods such as the Akaike information criterion (AIC), and Bayesian information criterion (BIC) have been employed for estimating the number of clusters (Fraley and Raftery, 2002). In our application, we chose to use BIC as a starting point to identify an inflection point (where BIC starts to tip-off) to identify an initial number of clusters, then merge similar clusters in a more empirical way, resulting in the most sensible clustering of the prevalence responses.

5.3.6 Distributed Implementation

There are two important challenges of the distributed computing implementation of the clustering algorithm. The first challenge is the storage and retrieval of the data throughout the computation process. The size of the data can be too large to be stored and computed using only one computing node. Thus in our implementation, we partition the data onto multiple storage nodes and execute the algorithm on each subset of data in a Map Reduce fashion. In more complex cases where data are naturally collected and stored in a decentralized approach, communicating all the data onto one centralized location can be very expensive. More sophisticated design of distributed storage topologies are required, as outlined in [103].

A second challenge is in the distributed computation itself for making the EM algorithm more scalable. However, without the independence assumption, the rows of the data matrix are coupled, thus the likelihood function cannot be decomposed in a way that allows distribution of the computation of its maximization. To address this challenge, we decompose and transform the correlation structure, allowing for the implementation of both the distributed data storage/retrieval and the parallel computation. In the E Step, the estimation of i^{th} response's membership probability in cluster k only requires information from its immediate neighbors. The expected sufficient statistics for each observed response can be computed independently in blocks given a current estimate of the parameters. In the M Step, the data in each neighborhood are transformed assuming the correlation structure. The parameter estimation can then be written in closed form summation, which can be efficiently implemented in a Map Reduce fashion in parallel.

The algorithm was implemented in Julia, a high-performance dynamic programming language for numerical and distributed computing [121].

5.4 Results

In this section, we present the results for the clustering approach to study the burden of chronic conditions for Medicaid-enrolled children in the United States. We first compare the clustering results under the Nominal EM (under the independence assumption) and the Spatial EM algorithm, to motivate the need of the additional computational effort of modeling the spatial structure in the chronic condition prevalence data. We then show the superior performance in runtime utilizing distributed computing versus sequential computing. Last, we provide results on the overall clustering throughout the United States with inference on differences of the chronic condition burden across states and urbanicity levels.

5.4.1 Nominal vs Spatial Clustering

We study and compare the clusters of census tracts under the nominal EM algorithm and the Spatial EM algorithm. Both algorithms use a randomized membership initiation scheme; that is, each census tract was randomly assigned to a cluster and an initial estimation of mean and covariance were calculated thereafter. In this section, for illustration purposes, we choose the number of clusters to be three since it produces the most meaningful division of census tracts among other selections for the number of clusters. Details on the model selection can be found in section 5.4.3.

Although most of the health conditions have very weak correlation, between -0.1 and 0.1, there still exists some moderate correlation, especially in the group of mental health conditions. Therefore, the features are first transformed into orthogonal principal components using principal component analysis. By using PCA, we are assuming that the feature correlation structures are approximately the same among different clusters. We calculate the sample correlation matrix for the entire population, and for each of the clusters. The 95% confidence interval for the pairwise difference between the correlation matrix for the entire population and the correlation matrices for cluster 1, 2, and 3 are [-0.01,-0.003], [-

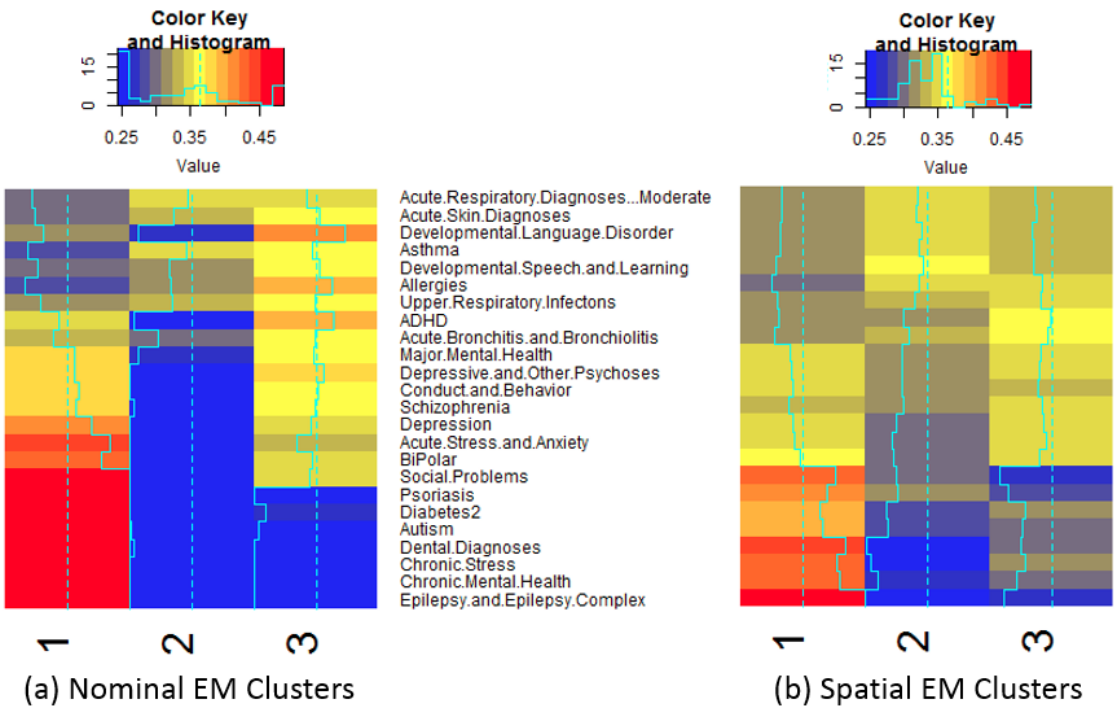


Figure 5.2: Heatmap of the prevalence in each cluster under (a) nominal EM Algorithm and (b) spatial EM Algorithm. The values are normalized so that each row sums to 1.

0.008,0.009], and [-0.008,0.003] respectively. The differences are minimal, which justifies using PCA in our study.

Figure 5.2 shows the prevalence for all the chronic conditions for each of the three clusters, contrasting the results based on the two clustering approaches. To better compare the composition of conditions across clusters, each row of the heatmap was normalized to sum to one. Under the Nominal EM algorithm, we see a clear separation of conditions in each cluster. Cluster 1 consists of 11,512 census tracts (17.7%), predominantly with chronic and moderate mental health diseases, along with some acute and major conditions. Cluster 2 consists of 25,473 census tracts (39.3%), where the prevalences of mental diseases are mostly low, with moderate prevalence in some respiratory and skin related diseases. Cluster 3 consists of 27,888 census tracts (43%), where moderate prevalences for all conditions exist, except for some severe chronic mental conditions, Diabetes, and Dental diseases. The clear separation is as expected, since the nominal EM algorithm clusters the census tracts

solely based on the absolute distribution of the prevalence of each condition.

Under the Spatial EM algorithm, the conditions are more blended in each cluster. Cluster 1 is very similar to the first cluster under the nominal EM algorithm in composition, with 23,896 census tracts (36.8%). Cluster 2 consists of 11,964 (18.4%) census tracts, where all of the respiratory related conditions, such as acute/moderate respiratory diagnoses, Asthma, Allergies, upper respiratory infections, Bronchiolitis among others are moderately prevalent. Cluster 3 consists of 29,013 census tracts (44.7%), with less respiratory and skin conditions, but more mental health conditions, in contrast to the Cluster 3 from the nominal EM algorithm.

Figure 5.3 shows the map of census tracts located in the states near the east coast of the United States. The census tracts are color coded based on the cluster membership under the two EM algorithms. Figure 5.4 takes a closer look at the areas near major cities, the coast line near Miami, New York City area, and Washington D.C.-Baltimore area. Generally, the locations of different clusters are similar, with rural areas consisting primarily of census tracts in Cluster 3, representing a larger portion of acute and major mental health issues. Pennsylvania, Vermont, New Hampshire and Maine exhibit considerably less prevalence for acute and major mental conditions.

The biggest difference between the two maps are the areas around major cities, labeled as black dots. The clusters generated from the nominal EM are homogeneous across the map. Census tracts in the same area tend to be from the same cluster, such as the coast line near Miami, and around New York city. The nominal EM algorithm failed to capture the heterogeneity in small areas, especially where the population is dense and diverse. On the contrary, in addition to the absolute distribution of the features, the Spatial EM algorithm accounts for the magnitude of prevalence values on the relative scale by modeling the spatial correlation in small areas. Therefore, it discovers relative differences on the spatial domain.

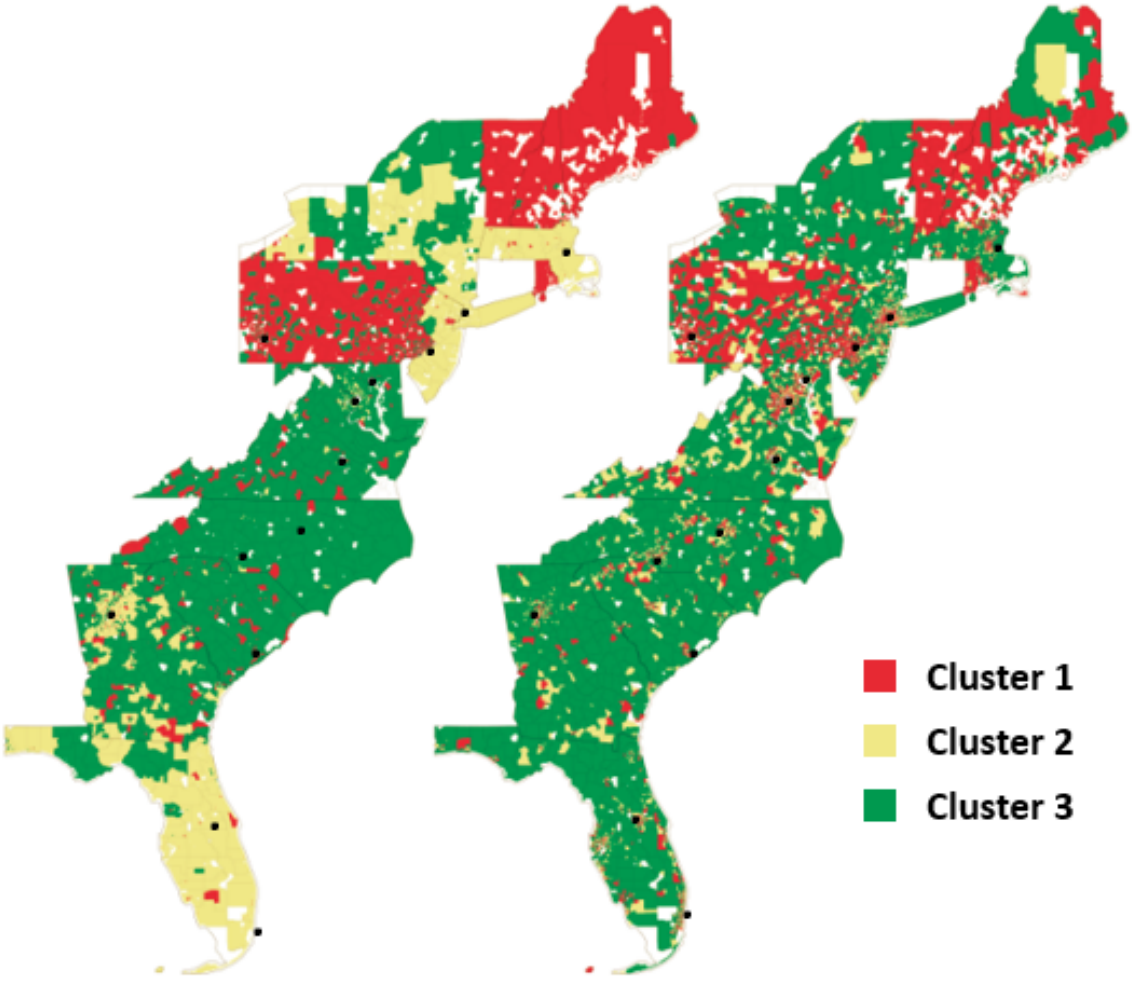


Figure 5.3: Maps of the census tracts located in the east coast states of the United States, color coded by the cluster membership under (a) nominal EM Algorithm and (b) spatial EM Algorithm. Each black dot represents a major city.

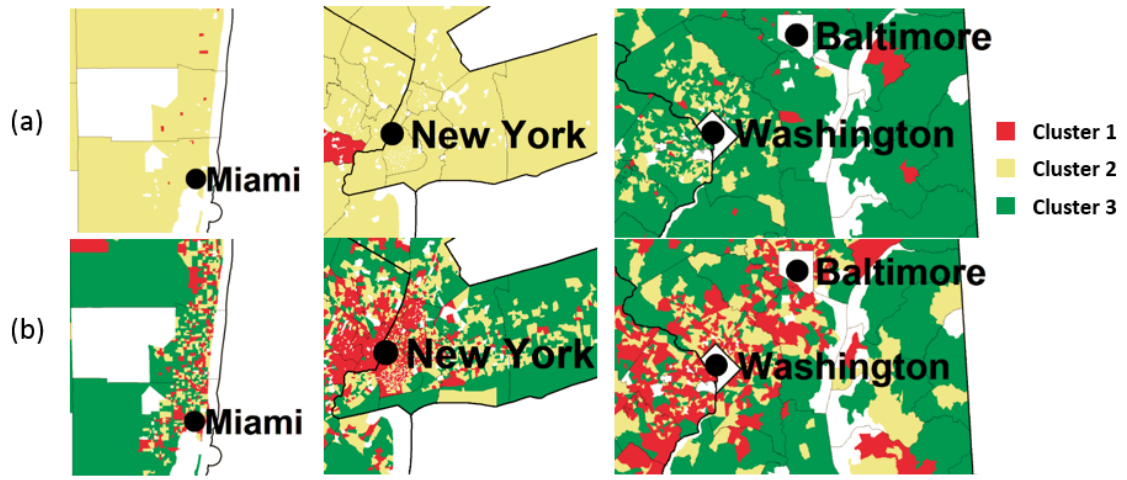


Figure 5.4: Zoomed-in maps of the census tracts close to major cities, color coded by the cluster membership under (a) nominal EM Algorithm and (b) spatial EM Algorithm.

5.4.2 Distributed Computation

The computation of the Spatial EM algorithm is significantly more complex than the nominal EM algorithm and requires more time and computing resources to execute. In this section, we illustrate how distributed computation can help alleviate the computational burden. In order to compare the computational results under the sequential and parallel implementations, we fix the number of iterations to be 100. Figure 5.5 shows the computational results with different number of computing cores (Intel Core Haswell Processors). The algorithm was written in Julia, and executed on a Linux server with X86-64 bit architecture.

The job execution required a total of 36.3 GB in memory allocation, thus infeasible to store and retrieve the data on a single machine/computing node – even the implementation using serial computation (one computing core) had to utilize a distributed storage framework. Running the algorithm using one computing core took more than 11 hours. This number can easily skyrocket to weeks as additional runs are required for sensitivity analysis, parameter tuning (e.g. number of clusters, size of neighborhood), and statistical inference, for example. Running the algorithm in a parallel fashion greatly reduces the computational time. With 10 computing cores, the run time was reduced to 1.8 hours, with

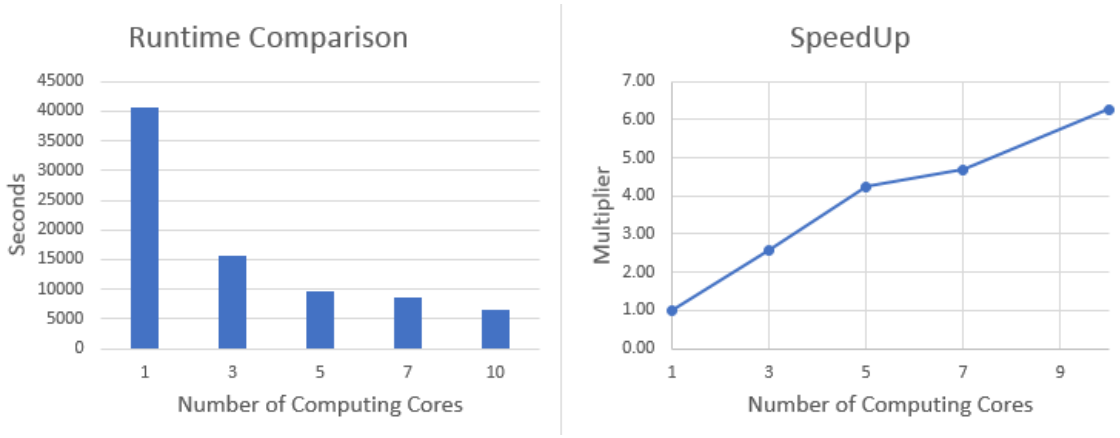
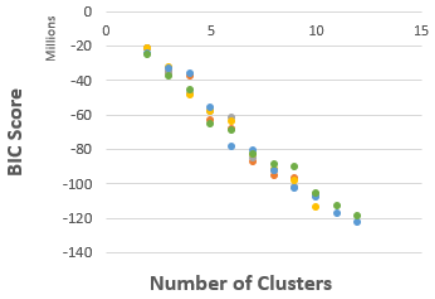


Figure 5.5: Runtime comparison in seconds and speed up with varying number of computing cores.

a 6.3 times speed up. We note that the speed up is not exactly proportional to the number of computing cores. In fact, the run time improvement is most significant with the first few added cores, and gradually decays as the number of cores further increases. This is commonly known as the Amdahl's Law, where the potential program speedup is defined by the fraction of code that can be parallelized [122]. In addition, other architectural and synchronization constraints such as memory-CPU bus bandwidth, communication bandwidth, load balancing and memory locks play key roles in coordinating the distributed execution and become more complex as the number of cores increases.

5.4.3 Model Selection

We use the BIC score as the model selection criteria to identify the number of clusters. In addition, since the clustering results tend to vary with different initiatializations, we run the algorithm five times for each setting to study the sensitivity of the imputed cluster memmbership and number of clusters to initiailization. Part (a) of Figure 5.6 shows that the BIC curve decreases with the number of clusters ranging from 2 to 12, and starts to flatten after 10 clusters. This suggests that, using the BIC criterion, the number of clusters chosen can be large thus BIC may not provide an upper threshold for the number of clusters. This is a possible indication that there are a few outliers that do not belong to any given cluster.



(a)

Matching Percentage

Size of Neighborhood	0	2	4	6	8	10
0		0.11	0.03	0.03	0.02	0.01
2	49%		0.18	0.15	0.1	0.07
4	43%	63%		0.13	0.2	0.16
6	44%	62%	71%		0.44	0.3
8	45%	62%	66%	81%		0.52
10	45%	62%	63%	75%	85%	

Adjusted Rand Index

(b)

Figure 5.6: (a) BIC score under different number of clusters. (b) The upper triangle shows the adjusted Rand index, and the lower triangle shows the matching percentage under varying neighborhood sizes.

We visually inspected the clustering results with varying number of clusters. As the number of clusters increases to more than three, additional clusters yield similar patterns, with a few very small clusters in size ($\leq$0.01%) that capture mostly the outlying features and big clusters that are not clearly distinguishable. Consequently, we choose to analyze the clustering with three clusters. More details of the analysis on the number of clusters can be found in Appendix B.

5.4.4 Sensitivity Analysis

To evaluate how the uncertainty or the sampling errors in the prevalence estimates affects the clustering results, we simulated 20 samples of prevalence data from a binomial sampling model derived from the data on Medicaid-enrolled children with each of the conditions, and compare the resulting clusters with the baseline dataset. For the 20 comparisons, the 95% confidence interval for the adjusted Rand Index is [0.896, 0.903], and the percentages of census tracts that changed membership are consistently less than 4%. Thus the standard errors from the prevalence estimates have limited impact on the clustering membership. This is due to the fact that the total number of member months of all Medicaid-enrolled children for most of the census tracts is large and thus the errors are small.

In order to reduce the computation complexity, we assumed a fixed neighborhood struc-

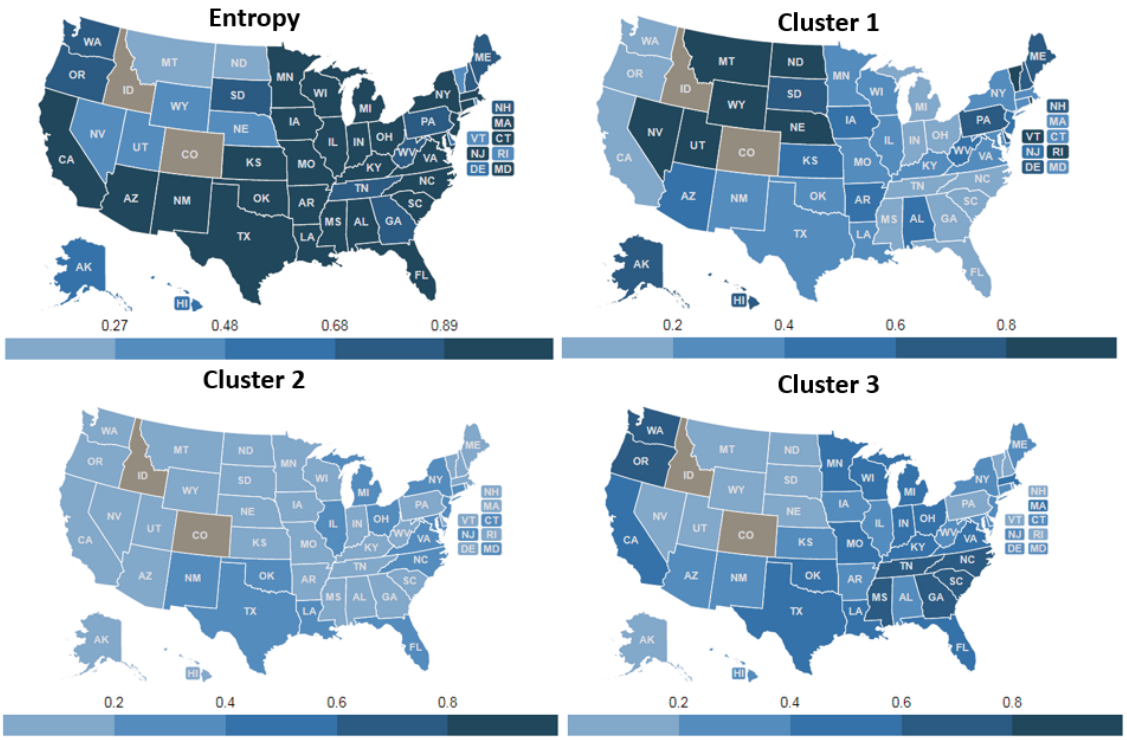


Figure 5.7: The entropy and proportional of census tracts within each state that belongs to each of the clusters.

ture. Part (b) of Figure 5.6 shows two measures for comparing any two clusterings obtained for varying neighborhood sizes. The upper triangle shows the adjusted Rand Index, which measures the similarity between two clusterings, adjusting for the chance of grouping, and the lower triangle shows the matching percentage. The nominal EM algorithm coincides with the spatial EM algorithm when the size of neighborhood is 0. When the neighborhood size is small, a slight change of the neighborhood can have big impact on the clustering result. This is an indication that the results can be sensitive when the spatial effect is not properly incorporated. As the neighborhood size increases, the similarity between clusterings improves drastically. It is therefore not necessary to consider the spatial correlation between every possible pair of locations, since a neighborhood of size 10 can produce a sufficiently stable clustering result.

5.4.5 Clustering Results: United States

Figure 5.7 displays the similarity (or dissimilarity) in clustering at the state level using the entropy measure (upper left map) and using the percentages for the three clusters. The west states have less variability in the clustering (lower entropy) than south west states. West states either predominantly are in cluster 1 (primarily represented by more severe chronic conditions) or cluster 3 (represented by mental health and respiratory chronic conditions). Cluster 2 (primarily represented by respiratory conditions) has low representation in most of the states except for a few southern states (e.g. FL, LA, NM, and TX) and northern states (e.g. IL, NY, NJ, VA). These state-level differences point to pediatric chronic conditions the states might need to focus on for disease management as well as prevention of severe outcomes.

Figure 5.8 shows the composition of each cluster by state and urbanicity for a subset of states. Urbanicity is defined using the rural-urban commuting area (RUCA) codes, which classify U.S. counties using measures of population density, urbanization, and daily commuting. The code is a single digit (1-9) classification, grouping counties based on the population of their metro area or their proximity to an urban area [123]. We further grouped the 1-9 code into 3 major categories. Category 1, with an RUCA index 1, represents urbanized metropolitans areas; Category 2, with an RUCA index 2-6, represents smaller metropolitans and micropolitan areas; Category 3, with an RUCA index 7 and above, represents small towns and rural areas. The distribution of the census tracts across the three clusters in these areas varies by state. As the census tracts become more rural, the proportion of clusters 1 and 2 decreases drastically; that is, chronic mental conditions, Diabetes, Autism and Respiratory conditions are more prevalent in urban areas. For states with the least dense population, Cluster 1 dominates across different urbanicity, and Cluster 2 is mostly nonexistent. These states exhibit much less heterogeneity comparing to states with higher population density and larger metropolitan areas.

Figure 5.9 shows community-level cluster membership for Georgia. Most rural Geor-

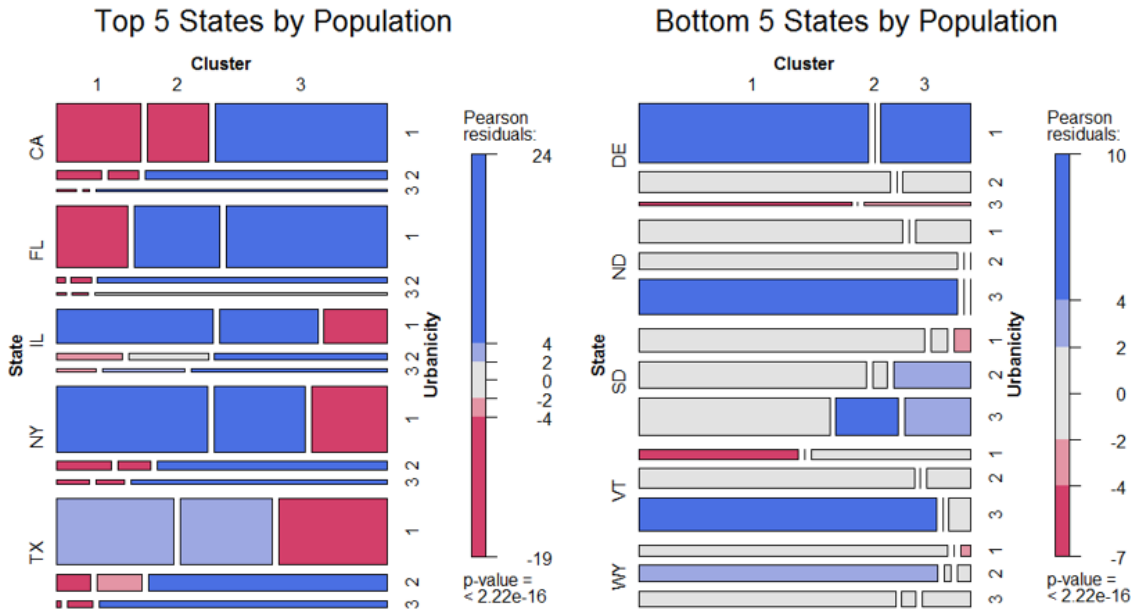


Figure 5.8: Visualization of the composition of each cluster by state and urbanicity for the top 5 and bottom 5 states by population under the Spatial EM Algorithm.

gia is predominantly in Cluster 3, with a mix of both mental health and respiratory chronic conditions, while suburban and urban areas are predominantly in Cluster 1 or 2, pointing to either heavily weighted mental and behavioral conditions or severe chronic conditions. We zoomed in the metropolitan Atlanta area, where several communities are assigned to Cluster 1 or 2. As noted in the heat map of Figure 5.2, the prevalences of the 25 conditions are differently weighted in Clusters 1 and 2; however, we see that there are many neighboring communities in the Atlanta area which are assigned to different clusters. Overall, this suggests that interventions for managing chronic conditions need to be much more targeted in urban areas.

Similar geographically granular analysis can be performed for other states. The maps for other states will be made available upon request from the authors of this paper.

5.5 Conclusions

The primary focus of this research paper is on deriving a spatial clustering of pediatric chronic conditions at the community level in the United States. The data supporting this

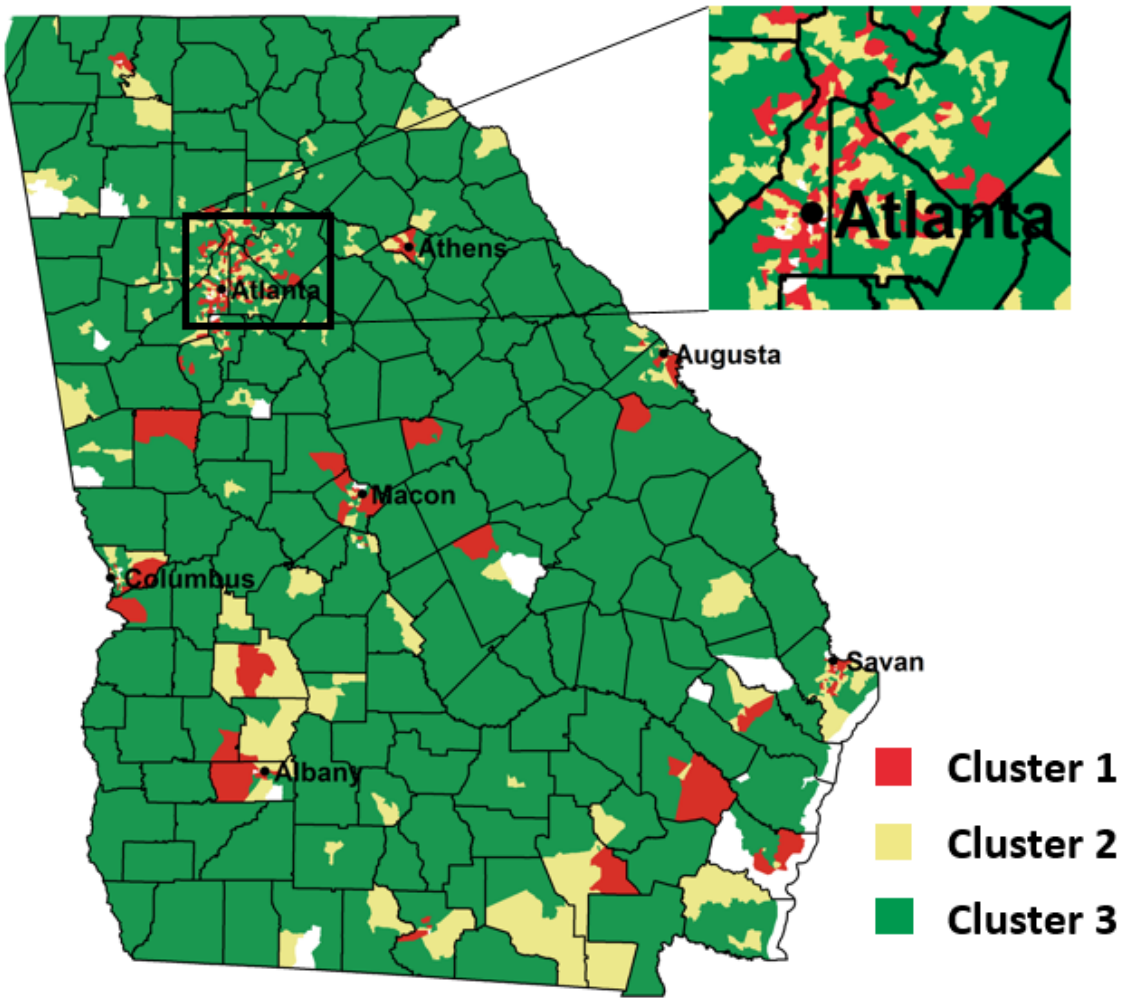


Figure 5.9: Clustering membership for the state of Georgia.

analysis consists of prevalences for 25 chronic conditions for the Medicaid-enrolled children.

The implementation of the spatial clustering approach relies on distributed computing to overcome the computational effort needed to perform the clustering analysis. While we were able to obtain the clustering after 11 hours of computing time with only the distribution of the data storage and retrieval, for a thorough analysis on the sensitivity of the clustering to the EM initialization and on the selection of the number of clusters, we needed much faster computations. Such large-scale studies can only be tackled by bridging statistical modeling and computational innovations.

This study has several limitations. The approach for estimating the prevalences at the census tract level from the prevalences observed at other geographic divisions, e.g. zip code, falls under the modifiable areal unit problem or MAUP . Our approach is one of the simplest MAUP approaches, with noted limitations [124]. However, obtaining more rigorous prevalence estimates at the census tract level requires extensive computational effort, which may be infeasible given the large scale of our data.

The correlation structure in the response data was assumed to follow a Matérn correlation function using the Euclidean distance between pairs of census tracts centroids. The distance metric can be further improved, such as to use road distance between centroids, or similarity measures in urbanicity, socio-economics factors, or demographics. Follow up analysis based on the clustering results and additional area specific covariates can provide insights in determining the main drivers of the spatial variation and discrepancies in prevalence. Although we limited the implementation of the proposed distributed model-based clustering analysis to spatial correlation, the proposed algorithm can be applied to any type of correlation structures. In addition, we assumed the correlation structure to be fixed for each feature and each of the C components. Alternatively, we can extend the model to concurrently re-evaluate the correlation functions for each feature and cluster component at each EM iteration as the membership changes. Moreover, the neighborhood size was

assumed to be fixed across all census tracts, which can be improved by a more granular definition of neighborhood based on urbanicity, for example.

Even though this research has several limitations, it has some important implications for interventions in managing chronic conditions. Many rural communities across the United States do not show a high burden of any particular condition, with similar weighting across respiratory conditions and behavioral & mental health conditions, with the lowest weight on more severe chronic conditions. This similarity in clustering across most of the rural communities points to that generally rural communities are in need of similar interventions, for example, improving access to mental and behavioral health providers. On the other hand, urban communities and some suburban communities present wide heterogeneity in clustering, with many of the urban communities being assigned in either high prevalence of severe chronic conditions or high prevalence of mental & behavioral conditions, which often are more severe for the Medicaid child population, overall pointing to a higher burden of severe conditions in some communities. While we cannot pinpoint the factors triggering such variations, we do recommend more targeted interventions for urban communities, with a focus on managing severe conditions.

Appendices

APPENDIX A

DERIVATION OF UTILIZATION SEQUENCES

This Appendix provides additional information on the derivation of the utilization sequences. A utilization sequence for an asthma-diagnosed child is derived based on the following set of information available in the Centers for Medicare and Medicaid (CMS) Medicaid Analytical Extract (MAX) medical claims data. The MAX claims are structured into inpatient care (IP), long-term care (LT), other care including outpatient services (OT), patient summary (PS) and prescription claim summary (RX) files, with a different set of files for each year and state. We merged the files for each state across all years to capture the longitudinal claims data for each patient in the study population using the Medicaid Statistical Information System (MSIS) identification number of each patient.

The data elements extracted from the MAX files are:

1. *Primary and secondary asthma diagnosis ICD-9 codes*: We extract only those claims with the following asthma-related ICD-9 codes: 493.00, 493.01, 493.02, 493.10, 493.11, 493.12, 493.20, 493.21, 493.22, 493.81, 493.82, 493.90, 493.91, 493.92. These are the only diagnosis codes corresponding to an asthma diagnosis available in the MAX files.

2. *Type of service (TOS) and Place of Service (POS) codes*: Both codes are available for each claim from the IP and OT files of the MAX extract. We use these codes to derive a provider type for each medical visit as shown in Table A.1. We consider multiple non-medication claims in the same day for the same patient to be one visit. In rare cases where there is more than one type of events during the same day, we use the first event of the day except in the case of an ER or HO event. If an ER or HO occurred, we prioritize ER and HO and set them as the event of the day.

TOS Code	*Logic*	*POS Code*	*Prov. Type*
		50: Federally qualified health center;	
12: Clinic	*OR*	*71: State or local public health clinic;*	*PO*
08: Physicians	*AND*	*11: Office*	*PO*
		72: Rural health clinic	
11: Outpatient hospital	*OR*	*23: Emergency room*	*ER*
01: Inpatient hospital	*AND*	*Any*	*HO*

Table A.1: Provider Type Crosswalk

3. *National Drug Code of Long-term asthma control medications*: These medications are taken regularly to control chronic symptoms and prevent asthma attacks and can be found in the RX file of the MAX extract. The medication types include: Inhaled Corticosteroids, Long-Acting Beta-Agonists (LABAs), Cromolyn and Tehophylline, Leukotriene Modifiers and Immunomodulators. We consider the claims for asthma control medication (ACM) and asthma short term medication (ASM) as event types in the stochastic network analysis.

APPENDIX B

MODEL SELECTION AND ESTIMATION

This Appendix takes the reader through a summary of the model selection algorithm and the model estimation portion of the algorithm. We provide derivations of the likelihood functions as well as the Kullback-Leibler distance between two Markov renewal processes (MRPs).

Model Selection

In this section of the appendix we describe the model selection algorithm. We seek to find the optimal clustering of sequences, given by $\vec{Z}_{\vec{R}}$, such that the BIC score is maximized. The BIC is an objective function that balances the tradeoff between maximizing the likelihood function while minimizing model size. For a model M,

$$BIC(M) = \ell(M) + |M| \cdot \log(R)/2,$$

where $\ell(M)$ is the log-likelihood of the model M, $|M|$ is the model size and R is the number of patients. Given the transition and interarrival parameters for the set of patients in profile k, P_k and Λ_k, for $k \in 1, \ldots, K$. For model M with K profiles, we will estimate $KS(S+1)-1$ parameters for the transition matrices, $P_k, k \in \{1, \ldots, K\}$, and KS^2 in the interarrival matrices, $\Lambda_k, k \in \{1, \ldots, K\}$.

In previous work the authors have used an EM algorithm to perform model estimation. However, such an algorithm requires the user to pre-specify the number of profiles, K, regardless of the number of true profiles. Additionally, each initialization may produce a different outcome, implying that a global optimum is not necessarily reached with each clustering result. However, with a satisfactory initialization the output will be nearly

optimal without complex calculation.

Other researchers favor a tree-based algorithm, where a distance metric is used to determine splits in the set of observation [125]. In contrast with the EM algorithm, the benefit of such a tree-based algorithm is that the the number of clusters can be determined after the clustering analysis is performed. However, it may not be guaranteed to maximize posterior likelihood of cluster membership. Therefore, we propose a joint tree-based, EM optimization algorithm that maximizes the BIC criterion.

As K and R increase, it becomes computationally intractable to consider all possible partitions to find the maximum BIC score. Therefore, we present an algorithm that searches for a nearly maximal BIC at each iteration. Our algorithm, as in [125], is guided by the Kullback-Leibler (KL) distance:

$$KL(Q_1||Q_1) = \int Q_1(x) \log \left(Q_1(x)/Q_2(x)\right) dx,$$

where Q_1 and Q_2 are the probability distributions under comparison. Specifically, we find the KL distance between the transition distribution out of each of the s_i for each individual sequence and the entire set of sequences in a given profile and then average across the $s_i, i \in \{1, \ldots, S\}$. (We provide the derivation of the KL distance in Appendix B.) We then order the average KL distances and find a nearly optimal partition in the observations to use as the initialization of the EM algorithm to maximize the posterior likelihood function. An overview of the algorithm is given below:

1. We begin with the null assumption, H_0, that all patients in a set belong to one profile. Find the population MLEs, $\bar{\Lambda}_{ij}$, and the transition matrix $\bar{P}_{ij}$ under the null hypothesis. Calculate the BIC_0 value.

2. Calculate the average KL distances between individual sequences and the one profile (null hypothesis), $D_{ave}(P, \Lambda||\bar{P}, \bar{\Lambda})$.

3. For a sufficiently large, equally-spaced set of the ordered average KL distances, (say,

50), $D_{(i)}$, let $W^{-}_{D_{(i)}}$ be the set of patients with average KL distances from the null distribution less than $D_{(i)}$, and $W^{+}_{D_{(i)}}$ be the set of patient with average KL distances from the null distribution greater than $D_{(i)}$. For each partition, $\{W^{-}_{D_{(i)}}, W^{+}_{D_{(i)}}\}$, calculate the BIC_A corresponding to the BIC value of the alternative hypothesis, H_A, that the set of sequences should be partitioned into two profiles.

4. Consider the partition $\{W^{*-}_{D_{(i)}}, W^{*+}_{D_{(i)}}\}$, such that the BIC score is maximized. Let this partition be the initialization for the EM algorithm. Recalculate the BIC score, call it BIC^{*}_{A} after the iterations of the EM algorithm.

5. If $BIC^{*}_{A} > BIC_0$, then divide the sequences into distinct profiles. Repeat steps (1)-(4) until no more divisions are made.

Likelihood Function Derivations

In this first subsection we provide derivations for the likelihood functions used in the model selection algorithm.

Derivation of Markov Chain Likelihood

Consider a discrete time Markov chain (DTMC) with a sequence of events denoted by $\vec{X}_L = (X_1, \ldots, X_L)$. The derivation of the likelihood function is given below:

$$\begin{aligned}
L(P|\vec{X}_L) &= \Pr(\vec{X}_L = \vec{s}_L) \\
&= \Pr(X_L = s_{i_L}|\vec{X}_{L-1} = \vec{s}_{L-1}) \\
&\times \Pr(X_{L-1} = s_{i_{L-1}}|\vec{X}_{L-2} = \vec{s}_{L-2}) \\
&\times \cdots \times P(X_2 = s_{i_2}|X_1 = s_{i_1}) \times P(X_1 = s_{i_1}) \\
&= \Pr(X_L = s_{i_L}|X_{L-1} = s_{i_{L-1}}) \\
&\times \Pr(X_{L-1} = s_{i_{L-1}}|X_{L-2} = s_{i_{L-2}}) \\
&\times \cdots \times \Pr(X_2 = s_{i_2}|X_1 = s_{i_1}) \times \Pr(X_1 = s_{i_1}) \\
&= P_{s_{i_{L-1}}, s_{i_L}} \times \cdots \times P_{s_{i_1}, s_{i_2}} \times P_{LC, s_{i_1}} \\
&= \prod_{l=1}^{L} P_{s_{i_{l-1}}, s_{i_l}} \times P_{LC, s_{i_1}}
\end{aligned}$$

Derivation of Markov Renewal Process Likelihood

The MRP is the continuous-time analog of a discrete-time Markov chain. The primary assumption of any Markov process is that it is 'memoryless', i.e. future states are only dependent on the current state of the system. Define $\tau_n = T_n - T_{n-1}$. Then we have that

$$\begin{aligned}
&Pr(\tau_{L+1} \leq t, X_{L+1} = s_j|X_1, T_1, \ldots, X_L, T_L) \\
&\quad Pr(\tau_{L+1} \leq t, X_{L+1} = s_j|X_L = s_i). \qquad \text{(B.1)}
\end{aligned}$$

Assuming that a patient sequence of events with timestamps follow a Markov renewal process, denoted by $(\vec{X}_L, \vec{T}_L)$, we provide the derivation of the likelihood function from

equation (B.1) below:

$$
\begin{aligned}
& L(P, \Lambda | \vec{X}_L, \vec{T}_L) \\
&= \Pr(\vec{X}_L = \vec{s}_L, \vec{T}_L = \vec{\tau}_L) \\
&= P(X_L = s_{i_L}, T_L = \tau_L | \vec{X}_{L-1} = \vec{s}_{L-1}, \vec{T}_{L-1} = \vec{\tau}_{L-1}) \\
&\times \cdots \times P(X_2 = s_{i_2}, T_2 = \tau_2 | X_1 = s_{i_1}, T_1 = \tau_1) \\
&\times P(X_1 = s_{i_1}, T_1 = \tau_1) \\
&= \Pr(X_L = s_{i_L}, T_L = \tau_L | X_{L-1} = s_{i_{L-1}}) \\
&\times \Pr(X_{L-1} = s_{i_{L-1}}, T_{L-1} = \tau_{L-1} | X_{L-2} = s_{i_{L-2}}) \\
&\times \cdots \times \Pr(X_2 = s_{i_2}, T_2 = \tau_2 | X_1 = s_{i_1}) \times \Pr(X_1 = s_{i_1})
\end{aligned}
$$

Next we make use of the following conditional probability rule:

$$
\begin{aligned}
& \Pr(X_l = s_{i_l}, T_l = \tau_l | X_{l-1} = s_{i_{l-1}}) \\
&= \Pr(T_l = \tau_l | X_{l-1} = s_{i_{l-1}}, X_l = s_{i_l}) \\
&\times \Pr(X_l = s_{i_l} | X_{l-1} = s_{i_{l-1}}).
\end{aligned}
$$

Combining the two previous equations completes the derivation.

Derivation of the KL Distance

Step (2) of the algorithm from the Model Selection section requires the calculation of the KL distance between the estimated one-step transition distributions of each patient sequence and the overall population. Let $\bar{P}$ be the transition matrix corresponding to profile k, and P be the transition matrix of observation r belonging to profile k. Likewise, let $\bar{\Lambda}$ contain the MLEs for the exponentialy distributed interarrival times for profile k, and Λ contain the MLEs for observation r belonging to profile k. Let $\bar{P}_{i,j}$, $P_{i,j}$, $\bar{\Lambda}_{i,j}$, and $\Lambda_{i,j}$ denote the transition probabilities and expected interarrival times between states s_i and s_j. We

can now derive a closed-form solution of the average KL distance between the transition distributions out of state s_i for an individual and a population.

Consider a utilization sequence such that $X_l = s_i$ at time t. We want to compare the probability of transition at time $T + \tau$ to state s_j of the patient to that of the all patients within the profile, where T was the last arrival time. This distribution is a finite mixture of exponential distributions: given that the next event is state s_j occuring with probability P_{ij}, the interarrival time is given by $Exp(\lambda_{ij})$. Using the P and Λ matrices we derive the KL distance between the individual and cluster distributions:

$$
\begin{aligned}
& d(P, \Lambda || \bar{P}, \bar{\Lambda}) \\
&= \sum_j \int_0^\infty P_{i,j}\lambda_{i,j} \exp\{-\lambda_{i,j}\tau\} \\
&\cdot \log\left(\frac{P_{i,j}\lambda_{i,j}\exp\{-\lambda_{i,j}\tau\}}{\bar{P}_{i,j}\bar{\lambda}_{i,j}\exp\{-\bar{\lambda}_{i,j}\tau\}}\right) d\tau \\
&= \sum_j P_{i,j} \int_0^\infty \lambda_{i,j}\exp\{-\lambda_{i,j}\tau\} \\
&\cdot \left[\log\left(\frac{P_{i,j}\lambda_{i,j}}{\bar{P}_{i,j}\bar{\lambda}_{i,j}}\right) + \log\left(\exp\{\bar{\lambda}_{i,j}\tau - \lambda_{i,j}\tau\}\right)\right] d\tau \\
&= \sum_j P_{i,j}\log\left(P_{i,j}/\bar{P}_{i,j}\right) + \log\left(\lambda_{i,j}/\bar{\lambda}_{i,j}\right) \\
&\quad + \bar{\lambda}_{i,j}\int_0^\infty \tau\lambda_{i,j}\exp\{-\lambda_{i,j}\tau\}d\tau \\
&\quad - \lambda_{i,j}\int_0^\infty \tau\lambda_{i,j}\exp\{-\lambda_{i,j}\tau\}d\tau \\
&= \sum_j P_{i,j}\left[\log\left(P_{i,j}/\bar{P}_{i,j}\right) + \log\left(\lambda_{i,j}/\bar{\lambda}_{i,j}\right) + \bar{\lambda}_{i,j}/\lambda_{i,j} - 1\right]
\end{aligned}
$$

Finally, we want to average across all states $s_i, i \in \{1, \ldots, S\}$, so we use the measure:

$$
D_{ave}(P, \Lambda || \bar{P}, \bar{\Lambda}) = \frac{\sum_{i=1}^S d(P, \Lambda || \bar{P}, \bar{\Lambda})}{S}.
$$

APPENDIX C

UTILIZATION CLUSTERING RESULTS

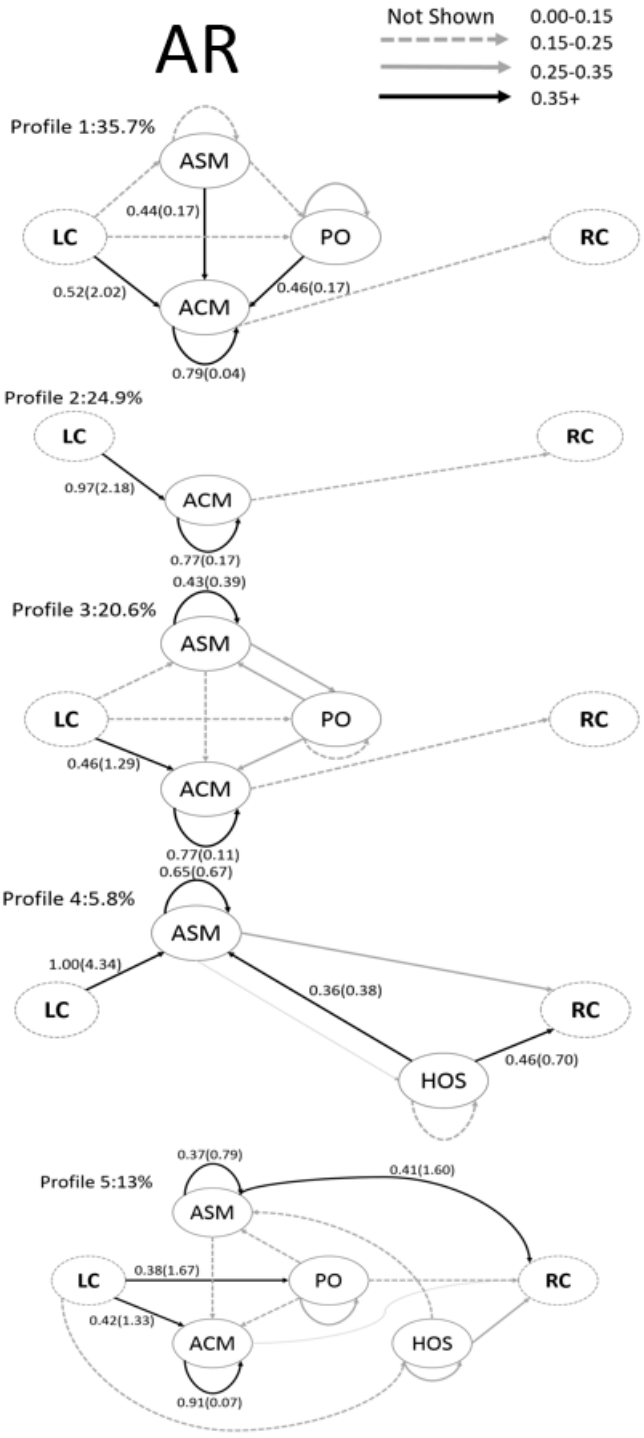


Figure C.1: Network graphs of etimated utilization profiles of AR. Transition probabilites are given on each edge along with the average interarrival times measured in months in parenthese.Some important edges with probability less than 0.15 are displayed in gray dotted lines.

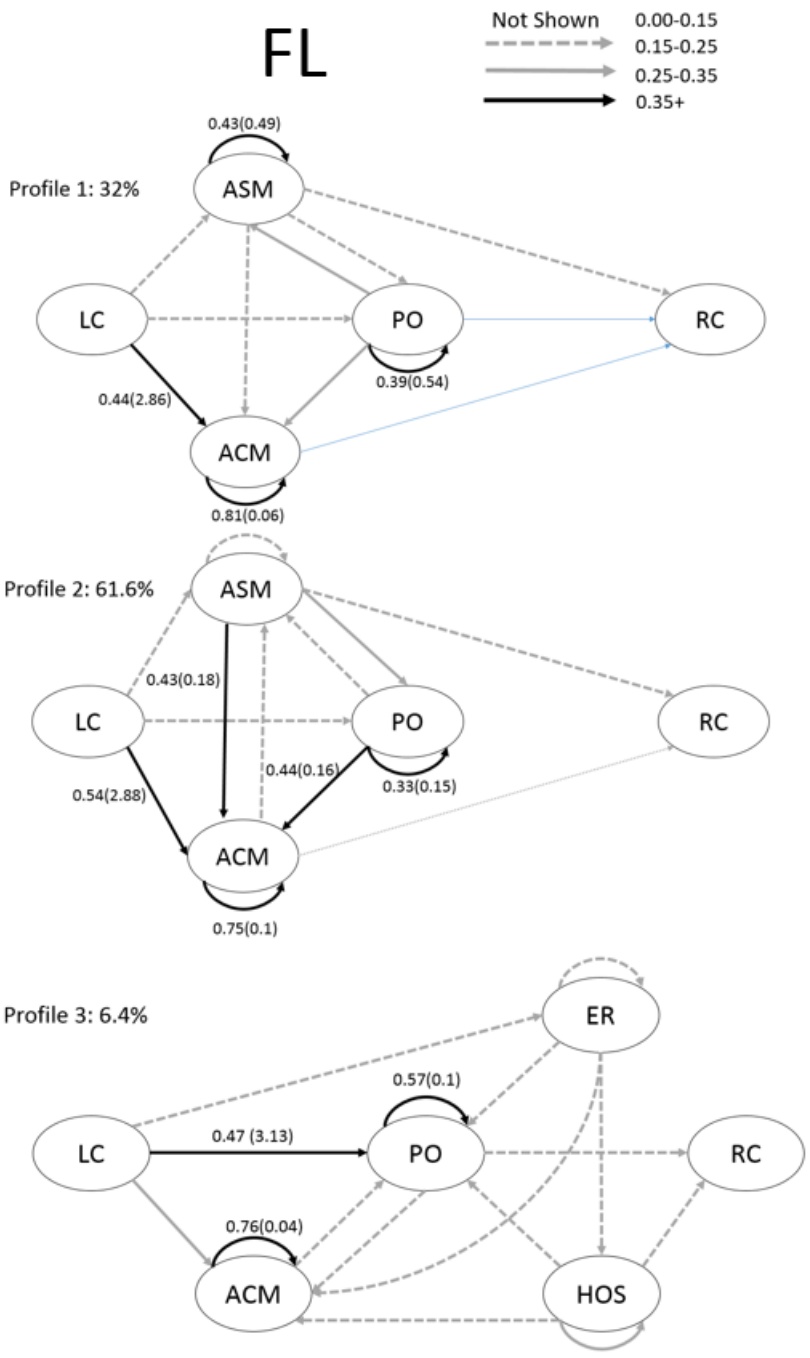


Figure C.2: Network graphs of etimated utilization profiles of FL. Transition probabilites are given on each edge along with the average interarrival times measured in months in parenthese.Some important edges with probability less than 0.15 are displayed in gray dotted lines.

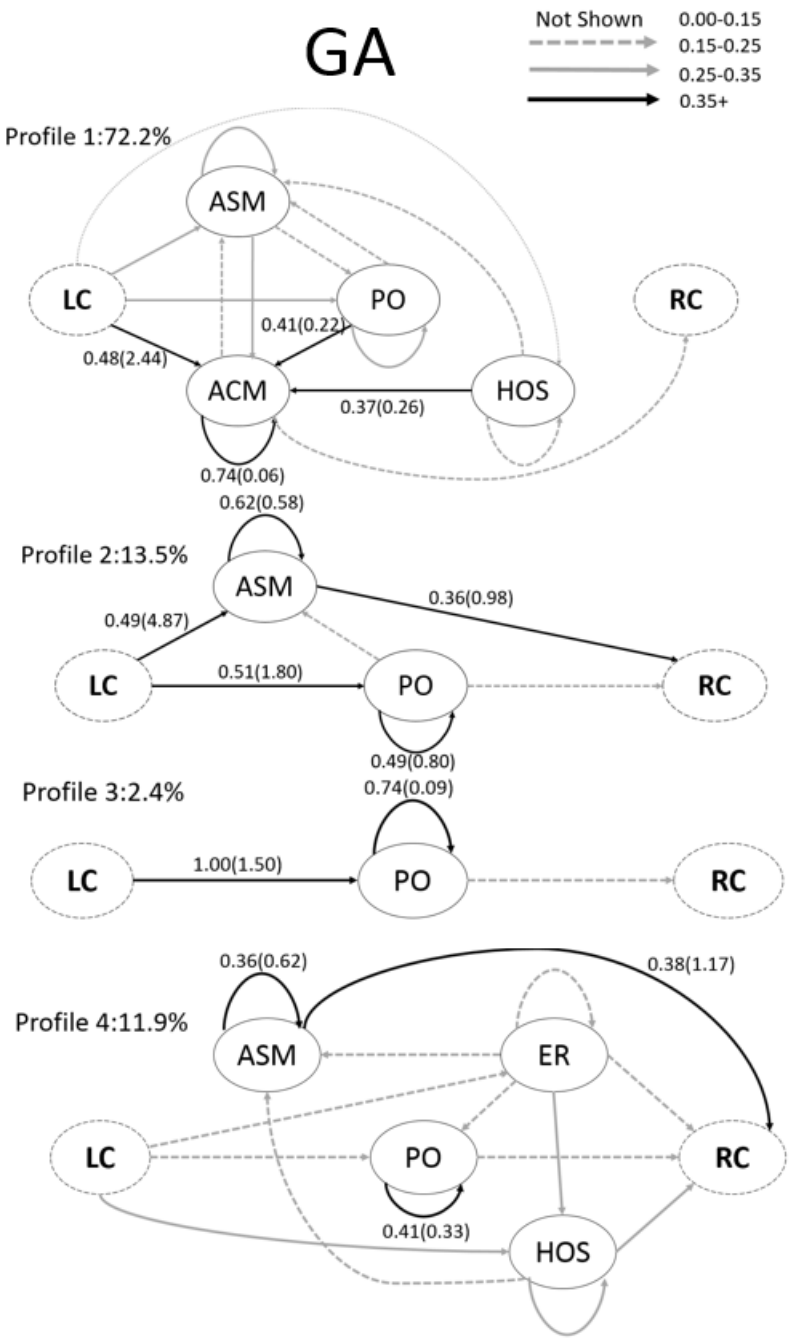


Figure C.3: Network graphs of etimated utilization profiles of GA. Transition probabilites are given on each edge along with the average interarrival times measured in months in parenthese.Some important edges with probability less than 0.15 are displayed in gray dotted lines.

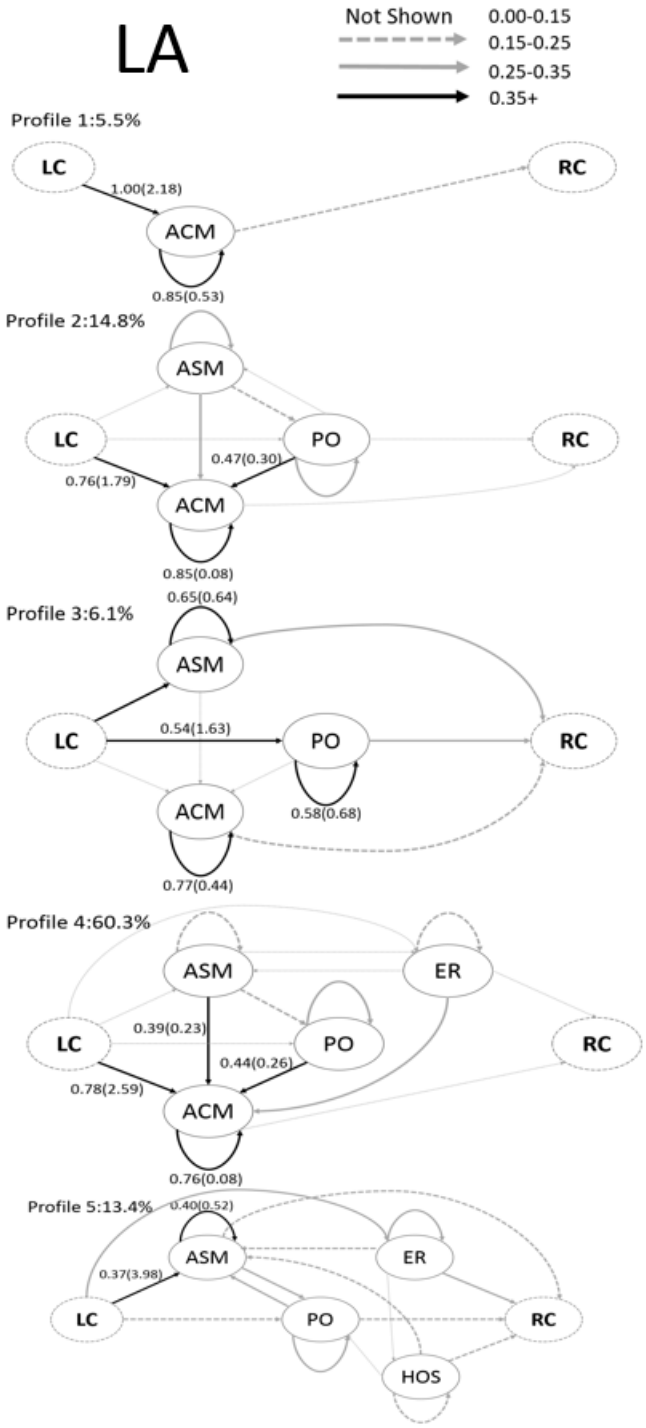


Figure C.4: Network graphs of etimated utilization profiles of LA. Transition probabilites are given on each edge along with the average interarrival times measured in months in parenthese.Some important edges with probability less than 0.15 are displayed in gray dotted lines.

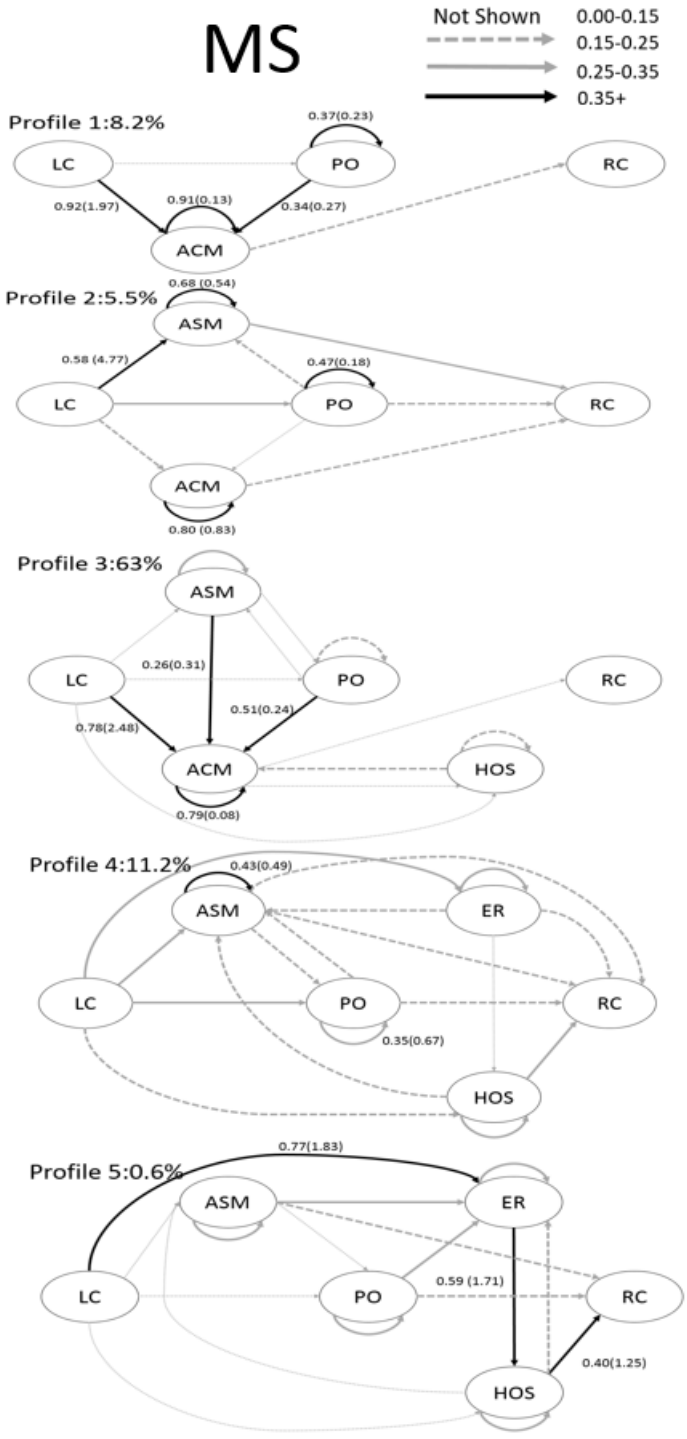


Figure C.5: Network graphs of etimated utilization profiles of MS. Transition probabilites are given on each edge along with the average interarrival times measured in months in parenthese.Some important edges with probability less than 0.15 are displayed in gray dotted lines.

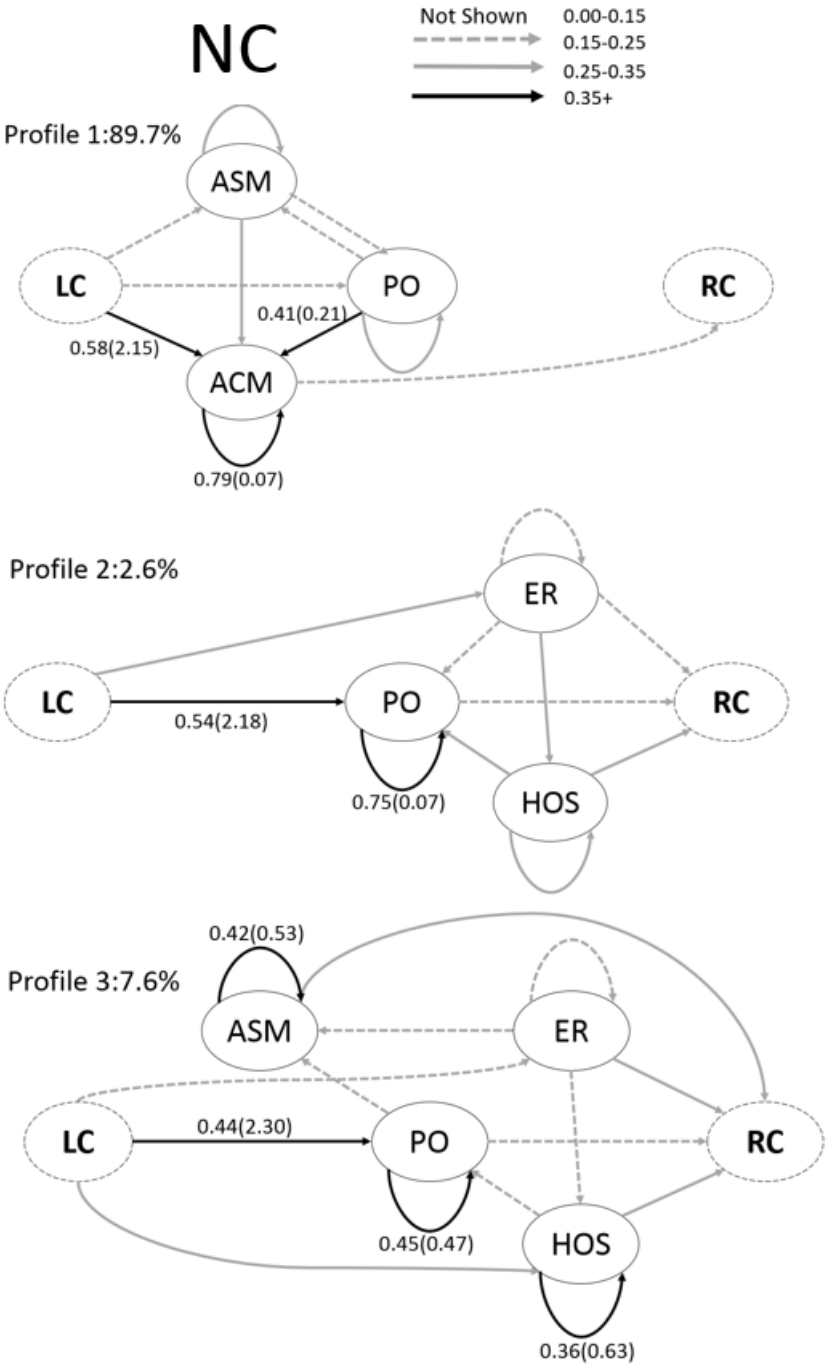


Figure C.6: Network graphs of etimated utilization profiles of NC. Transition probabilites are given on each edge along with the average interarrival times measured in months in parenthese.Some important edges with probability less than 0.15 are displayed in gray dotted lines.

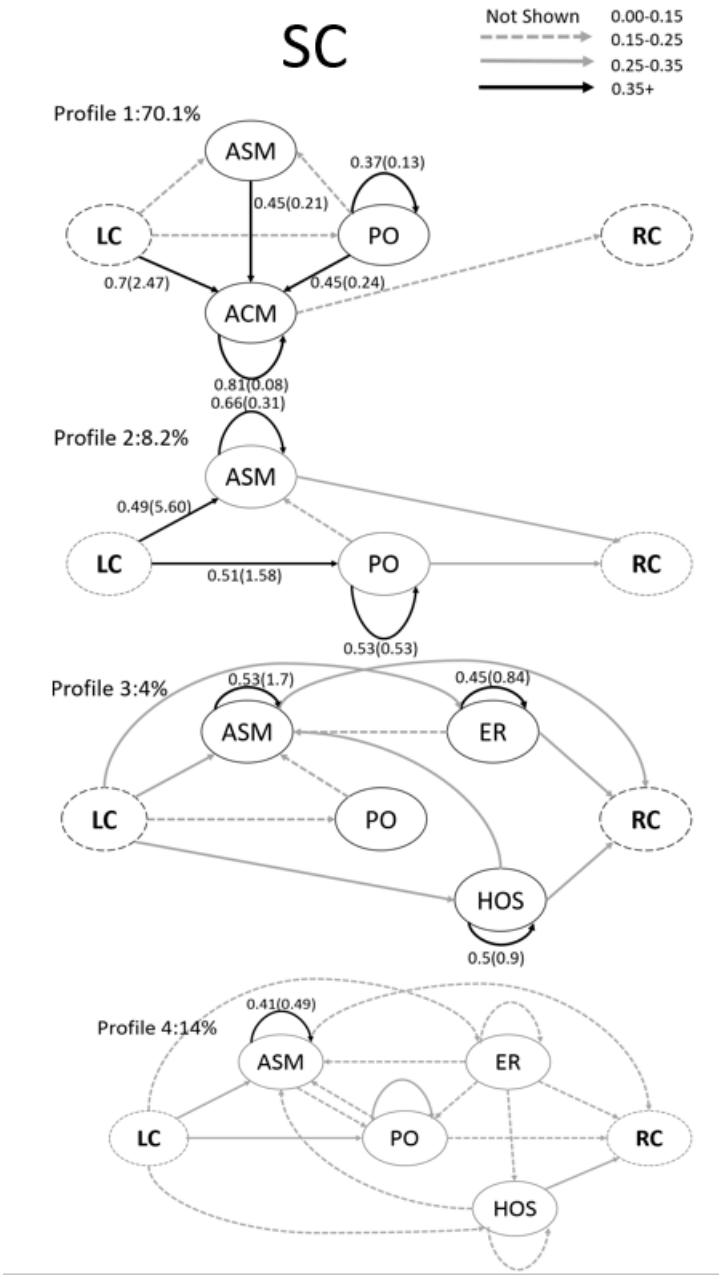


Figure C.7: Network graphs of etimated utilization profiles of SC. Transition probabilites are given on each edge along with the average interarrival times measured in months in parenthese.Some important edges with probability less than 0.15 are displayed in gray dotted lines.

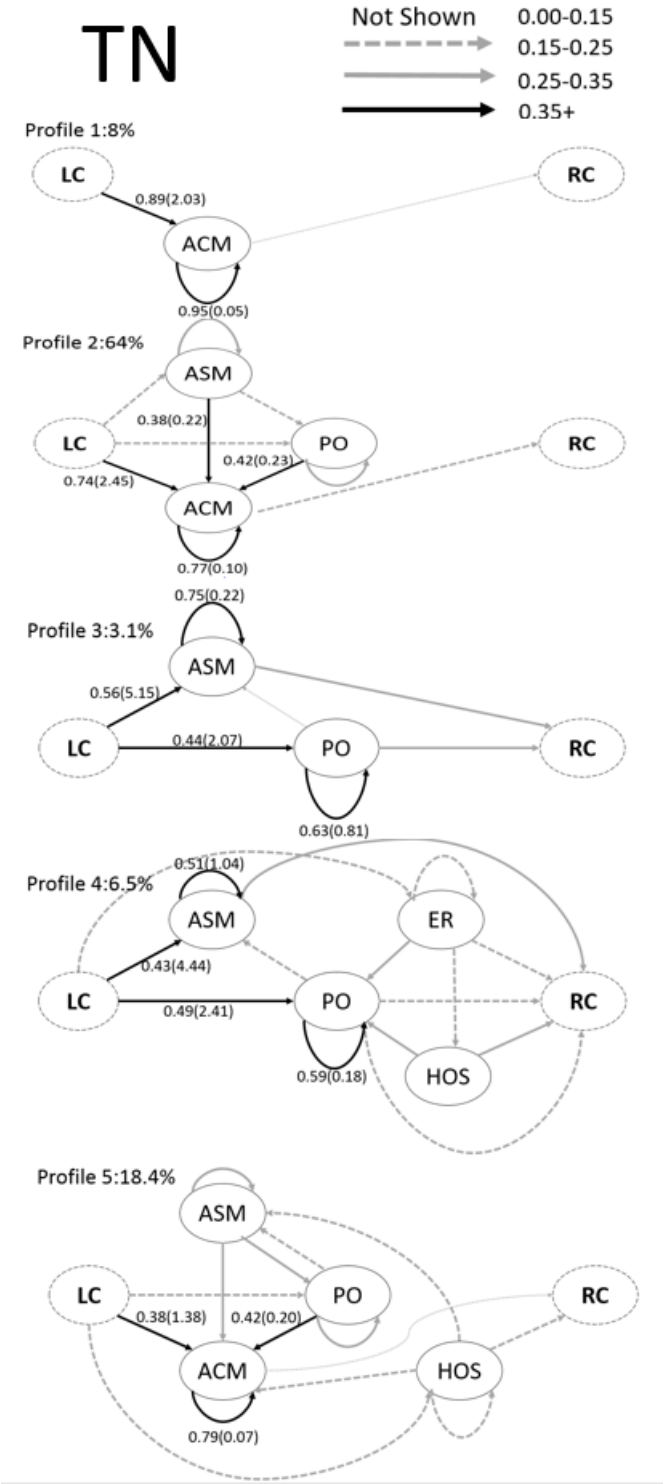


Figure C.8: Network graphs of etimated utilization profiles of TN. Transition probabilites are given on each edge along with the average interarrival times measured in months in parenthese.Some important edges with probability less than 0.15 are displayed in gray dotted lines.

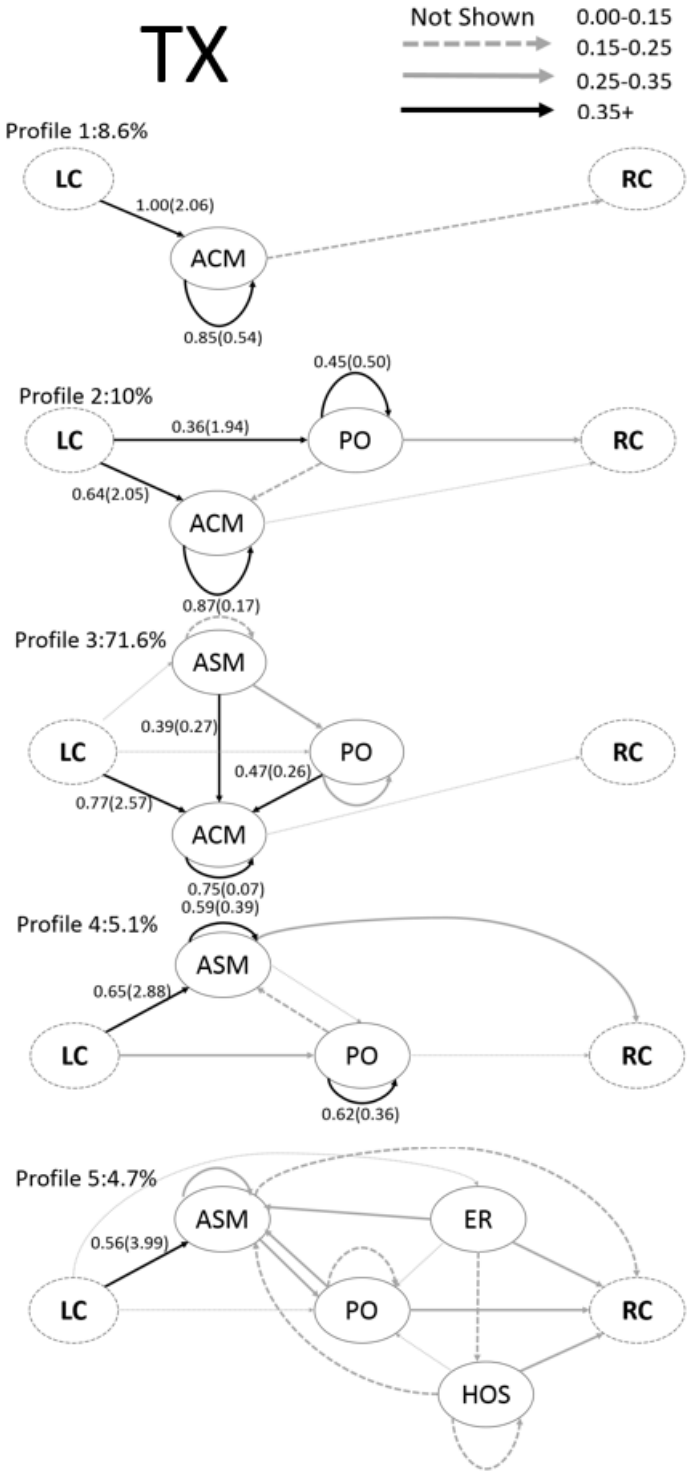


Figure C.9: Network graphs of etimated utilization profiles of TX. Transition probabilites are given on each edge along with the average interarrival times measured in months in parenthese.Some important edges with probability less than 0.15 are displayed in gray dotted lines.

REFERENCES

[1] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[2] L. Sweeney, A. Abu, and J. Winn, "Identifying participants in the personal genome project by name," 2013.

[3] J. Vaidya, B. Shafiq, X. Jiang, and L. Ohno-Machado, "Identifying inference attacks against healthcare data repositories," *AMIA Summits on Translational Science Proceedings*, vol. 2013, p. 262, 2013.

[4] M. L. Braunstein, *Health informatics in the cloud*. Springer, 2012.

[5] C. for Disease Prevention and Control, "Asthma: State data profiles," 2011.

[6] ——, "National health interview study raw data," 2011.

[7] L. B. Bacharier, R. C. Strunk, D. Mauger, D. White, R. F. Lemanske Jr, and C. A. Sorkness, "Classifying asthma severity in children: Mismatch between symptoms, medication use, and lung function," *American journal of respiratory and critical care medicine*, vol. 170, no. 4, pp. 426–432, 2004.

[8] D. J. Gottlieb, A. S. Beiser, and G. T. O'connor, "Poverty, race, and medication use are correlates of asthma hospitalization rates: A small area analysis in boston," *Chest*, vol. 108, no. 1, pp. 28–35, 1995.

[9] L. National Heart and B. Institute, *How is asthma treated and controlled?* 2015.

[10] P. Barnes, B Jonsson, and J. Klim, "The costs of asthma," *European Respiratory Journal*, vol. 9, no. 4, pp. 636–642, 1996.

[11] E. Bateman, *The economic burden of uncontrolled asthma across europe and the asia-pacific region: Can we afford to not control asthma?* 2006.

[12] E. Juniper, M. Wisniewski, F. Cox, A. Emmett, K. Nielsen, and P. O'Byrne, "Relationship between quality of life and clinical status in asthma: A factor analysis," *European Respiratory Journal*, vol. 23, no. 2, pp. 287–291, 2004.

[13] M. E. McGrady and K. A. Hommel, "Medication adherence and health care utilization in pediatric chronic illness: A systematic review," *Pediatrics*, vol. 132, no. 4, pp. 730–740, 2013.

[14] C. for Disease Prevention and Control, "The 6—18 initiative: Accelerating evidence into action," 2015.

[15] A. B. R. de Arellano and S. M. Wolfe, *Unsettling scores: A ranking of state Medicaid programs*. Public Citizen, 2007.

[16] W. S. Pearson, S. A. Goates, S. D. Harrykissoon, and S. A. Miller, "Peer reviewed: State-based medicaid costs for pediatric asthma emergency department visits," *Preventing chronic disease*, vol. 11, 2014.

[17] D. B. Wakefield and M. M. Cloutier, "Modifications to hedis and cste algorithms improve case recognition of pediatric asthma," *Pediatric pulmonology*, vol. 41, no. 10, pp. 962–971, 2006.

[18] N. C. for Quality Assurance, "Improving outcomes in asthma: Advancing quality using ncqa hedis measures.," 2011.

[19] N. H. Blood and L. Institute, "Expert panel report 3: Guidelines for the diagnosis and management of asthma," 2007.

[20] J. E. DeVoe, R. Gold, P. McIntire, J. Puro, S. Chauvie, and C. A. Gallia, "Electronic health records vs medicaid claims: Completeness of diabetes preventive care data in community health centers," *The Annals of Family Medicine*, vol. 9, no. 4, pp. 351–358, 2011.

[21] L. T. Piecoro, M. Potoski, J. C. Talbert, and D. E. Doherty, "Asthma prevalence, cost, and adherence with expert guidelines on the utilization of health care services and costs in a state medicaid population.," *Health services research*, vol. 36, no. 2, p. 357, 2001.

[22] V. L. Byrd, A. H. Dodd, *et al.*, "Assessing the usability of max 2008 encounter data for comprehensive managed care," *Medicare Care & Medicaid Research Review*, vol. 3, no. 1, E1–E19, 2013.

[23] ——, "Assessing the usability of encounter data for enrollees in comprehensive managed care 2010-2011," Mathematica Policy Research, Tech. Rep., 2015.

[24] M. Gentili, P. Harati, and N. Serban, "Projecting the impact of the affordable care act provisions on accessibility and availability of primary care providers for the adult population in georgia," *American Journal of Public Health*, vol. 106, no. 8, pp. 1470–1476, 2016.

[25] M. Gentili, K. Isett, N. Serban, and J. Swann, "Small-area estimation of spatial access to care and its implications for policy," *Journal of Urban Health-Bulletin of the New York Academy of Medicine*, vol. 92, no. 5, pp. 864–909, 2015.

[26] T. K. Dasaklis, C. P. Pappis, and N. P. Rachaniotis, "Epidemics control and logistics operations: A review," *International Journal of Production Economics*, vol. 139, no. 2, pp. 393–410, 2012.

[27] E. K. Lee, C.-H. Chen, F. Pietz, and B. Benecke, "Modeling and optimizing the public-health infrastructure for emergency response," *Interfaces*, vol. 39, no. 5, pp. 476–490, 2009.

[28] S. H. Owen and M. S. Daskin, "Strategic facility location: A review," *European Journal of Operational Research*, vol. 111, no. 3, pp. 423–447, 1998.

[29] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[30] T. Yamada, "A network flow approach to a city emergency evacuation planning," *International Journal of Systems Science*, vol. 27, no. 10, pp. 931–936, 1996.

[31] J. Heir Stamm, N. Serban, J. Swann, and P. Wortley, "Quantifying and explaining accessibility with application to the 2009 h1n1 vaccination campaign," *Health Care Management Science*, 2015.

[32] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 6.

[33] A. BenTal, L. E. Ghaoui, and A. Nemirovski, "Robust optimization," *Robust Optimization*, pp. 1–542, 2009.

[34] D. Bertsimas and M. Sim, "The price of robustness," *Operations Research*, vol. 52, no. 1, pp. 35–53, 2004.

[35] J. M. Mulvey, R. J. Vanderbei, and S. A. Zenios, "Robust optimization of large-scale systems," *Operations Research*, vol. 43, no. 2, pp. 264–281, 1995.

[36] Y. C. Jin and B. Sendhoff, "Trade-off between performance and robustness: An evolutionary multiobjective approach," *Evolutionary Multi-Criterion Optimization, Proceedings*, vol. 2632, pp. 237–251, 2003.

[37] S. C. H. Leung, S. O. S. Tsang, W. L. Ng, and Y. Wu, "A robust optimization model for multi-site production planning problem in an uncertain environment," *European Journal of Operational Research*, vol. 181, no. 1, pp. 224–238, 2007.

[38] S. P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[39] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends® in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[40] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for machine learning*. Mit Press, 2012, ISBN: 026201646X.

[41] S. J. Wright, "Accelerated block-coordinate relaxation for regularized optimization," *Siam Journal on Optimization*, vol. 22, no. 1, pp. 159–186, 2012.

[42] L Wasserman, *All of Nonparametric Statistics*. Springer, New York, 2006.

[43] H. Xu, C. Caramanis, and S. Mannor, "Statistical optimization in high dimensions," *Operations Research*, vol. 64, no. 4, pp. 958–979, 2016.

[44] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009, ISBN: 978-0387848570.

[45] J. Linderoth, A. Shapiro, and S. Wright, "The empirical behavior of sampling methods for stochastic programming," *Annals of Operations Research*, vol. 142, no. 1, pp. 215–241, 2006.

[46] S. Sen, R. D. Doverspike, and S. Cosares, "Network planning with random demand," *Telecommunication Systems*, vol. 3, no. 1, pp. 11–30, 1994.

[47] V. Pillac, M. Cebrian, and P. Van Hentenryck, "A column-generation approach for joint mobilization and evacuation planning," *Constraints*, vol. 20, no. 3, pp. 285–303, 2015.

[48] H. D. Sherali, T. B. Carter, and A. G. Hobeika, "A location-allocation model and algorithm for evacuation planning under hurricane flood conditions," *Transportation Research Part B-Methodological*, vol. 25, no. 6, pp. 439–452, 1991.

[49] J.-L. Wang, J.-M. Chiou, and H.-G. Muller, "Review of functional data analysis," *Annu. Rev. Statist.*, pp. 1–41, 2015.

[50] I. P. Androulakis, V. Visweswaran, and C. A. Floudas, "Distributed decomposition-based approaches in global optimization," in *State of the Art in Global Optimization: Computational Methods and Applications*. Boston, MA: Springer US, 1996, pp. 285–301, ISBN: 978-1-4613-3437-8.

[51] E. Camponogara and L. B. De Oliveira, "Distributed optimization for model predictive control of linear-dynamic networks," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 39, no. 6, pp. 1331–1338, 2009.

[52] G. Inalhan, D. M. Stipanovic, and C. J. Tomlin, "Decentralized optimization, with application to multiple aircraft coordination," in *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.*, vol. 1, 2002, 1147–1155 vol.1, ISBN: 0191-2216.

[53] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[54] D. P. Palomar and C. Mung, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, 2006.

[55] R. L. Raffard, C. J. Tomlin, and S. P. Boyd, "Distributed optimization for cooperative agents: Application to formation flight," in *Decision and Control, 2004. CDC. 43rd IEEE Conference on*, vol. 3, IEEE, 2004, pp. 2453–2459, ISBN: 0780386825.

[56] P. Richtárik and M. Takáč, "Parallel coordinate descent methods for big data optimization," *Mathematical Programming*, vol. 156, no. 1-2, pp. 433–484, 2016.

[57] Y. Shastri, A. Hansen, L. Rodríguez, and K. C. Ting, "A novel decomposition and distributed computing approach for the solution of large scale optimization models," *Computers and electronics in agriculture*, vol. 76, no. 1, pp. 69–79, 2011.

[58] A. Simonetto and H. Jamali-Rad, "Primal recovery from consensus-based dual decomposition for distributed convex optimization," *Journal of Optimization Theory and Applications*, vol. 168, no. 1, pp. 172–197, 2016.

[59] H. Terelius, U. Topcu, and R. M. Murray, "Decentralized multi-agent optimization via dual decomposition," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 11 245–11 251, 2011.

[60] L. Xiao, M. Johansson, and S. P. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *Ieee Transactions on Communications*, vol. 52, no. 7, pp. 1136–1144, 2004.

[61] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, 2012.

[62] N. Parikh and S. P. Boyd, "Block splitting for distributed optimization," *Mathematical Programming Computation*, vol. 6, no. 1, pp. 77–102, 2014.

[63] J. Guo, G. Hug, and O. Tonguz, "Intelligent partitioning in distributed optimization of electric power systems," *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1249–1258, 2016.

[64] J. Allison, M. Kokkolaras, and P. Papalambros, "Optimal partitioning and coordination decisions in decomposition-based design optimization," *Journal of Mechanical Design*, vol. 131, no. 8, p. 081 008, 2009.

[65] D. P. Bertsekas, *Nonlinear programming*. Belmont, Mass.: Athena Scientific, 1995, x, 646 p. ISBN: 1886529140.

[66] S. P Boyd, "Subgradient methods," *Lecture Notes*, 2014.

[67] J. Goffin, "On convergence rates of subgradient optimization methods," *Mathematical programming*, vol. 13, no. 1, pp. 329–347, 1977.

[68] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical review E*, vol. 69, no. 6, p. 066 133, 2004.

[69] ——, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[70] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3, pp. 75–174, 2010.

[71] T. Khaniyev, S. Elhedhli, and F. S. Erenay, "Structure detection in mixed integer programs," *INFORMS Journal on Computing*, 2017, accepted.

[72] A. Nedic and A. Ozdaglar, "Approximate primal solutions and rate analysis for dual subgradient methods," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1757–1780, 2009.

[73] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066 111, 2004.

[74] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026 113, 2004.

[75] J. Bezanson, S. Karpinski, V. B. Shah, and A. Edelman, "Julia: A fast dynamic language for technical computing," *ArXiv preprint arXiv:1209.5145*, 2012.

[76] M. Gentili, N. Serban, P. Harati, J. O'Connor, and J. Swann, "Quantifying disparities in accessibility and availability of pediatric primary care with implications for policy," *Health Services Research, (in press)*, 2017.

[77] D. P. Bertsekas, "Incremental gradient, subgradient, and proximal methods for convex optimization: A survey," *Optimization for Machine Learning*, vol. 2010, no. 1-38, p. 3, 2011.

[78] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical programming*, vol. 103, no. 1, pp. 127–152, 2005.

[79] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[80] R. D. Nowak, "Distributed em algorithms for density estimation and clustering in sensor networks," *IEEE transactions on signal processing*, vol. 51, no. 8, pp. 2245–2253, 2003.

[81] J. Wolfe, A. Haghighi, and D. Klein, "Fully distributed em for very large datasets," in *Proceedings of the 25th international conference on Machine learning*, ACM, 2008, pp. 1184–1191, ISBN: 1605582050.

[82] K. Knobe, J. D. Lukas, and G. L. Steele, "Data optimization: Allocation of arrays to reduce communication on simd machines," *Journal of parallel and Distributed Computing*, vol. 8, no. 2, pp. 102–118, 1990.

[83] J. Hromkovič, *Communication complexity and parallel computing*. Springer Science & Business Media, 2013.

[84] S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.

[85] Center for Medicare and Medicaid Services, *September 2017 medicaid and chip enrollment data highlights*, `https://www.medicaid.gov/medicaid/program-information/medicaid-and-chip-enrollment-data/report-highlights/index.html`, Online, 2017.

[86] ——, *Quality of care health disparities*, `https://www.medicaid.gov/medicaid/quality-of-care/improvement-initiatives/health-disparities/index.html`, Online, 2017.

[87] The World Health Organization, *Chronic diseases and their common risk factors*, `http://www.who.int/chp/chronic_disease_report/media/Factsheet1.pdf`, Online, 2005.

[88] W. C. Cockerham, B. W. Hamby, and G. R. Oates, *The social determinants of chronic disease*, 2017.

[89] A. Lawson, A. Biggeri, E Lessaffre, *et al.*, "Disease mapping and risk assessment for public health," 1999.

[90] P. Elliot, J. C. Wakefield, N. G. Best, D. J. Briggs, *et al.*, *Spatial epidemiology: Methods and applications.* Oxford University Press, 2000.

[91] L. A. Waller and C. A. Gotway, *Applied spatial statistics for public health data*. John Wiley & Sons, 2004, vol. 368.

[92] J. Wakefield, "Disease mapping and spatial regression with count data," *Biostatistics*, vol. 8, no. 2, pp. 158–183, 2006.

[93] S. Openshaw, M. Charlton, C. Wymer, and A. Craft, "A mark 1 geographical analysis machine for the automated analysis of point data sets," *International Journal of Geographical Information System*, vol. 1, no. 4, pp. 335–358, 1987.

[94] J. Besag and J. Newell, "The detection of clusters in rare diseases," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp. 143–155, 1991.

[95] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, 1996, pp. 226–231.

[96] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.

[97] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data," *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[98] M. Wang, A. Wang, and A. Li, "Mining spatial-temporal clusters from geo-databases," *Advanced Data Mining and Applications*, pp. 263–270, 2006.

[99] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1026–1038, 2002.

[100] H. Jiang and N. Serban, "Clustering random curves under spatial interdependence with application to service accessibility," *Technometrics*, vol. 54, no. 2, pp. 108–119, 2012.

[101] P. J. Green and S. Richardson, "Hidden markov models and disease mapping," *Journal of the American statistical association*, vol. 97, no. 460, pp. 1055–1070, 2002.

[102] C.-T. Chu, S. K. Kim, Y.-A. Lin, Y. Yu, G. Bradski, K. Olukotun, and A. Y. Ng, "Map-reduce for machine learning on multicore," in *Advances in neural information processing systems*, 2007, pp. 281–288.

[103] J. Wolfe, A. Haghighi, and D. Klein, "Fully distributed em for very large datasets," in *Proceedings of the 25th international conference on Machine learning*, ACM, 2008, pp. 1184–1191.

[104] J. M. Neff, V. L. Sharp, J. Muldoon, J. Graham, J. Popalisky, and J. C. Gay, "Identifying and classifying children with chronic conditions using administrative data with the clinical risk group classification system," *Ambulatory Pediatrics*, vol. 2, no. 1, pp. 71–79, 2002.

[105] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[106] C. Fraley and A. E. Raftery, "How many clusters? which clustering method? answers via model-based cluster analysis," *The computer journal*, vol. 41, no. 8, pp. 578–588, 1998.

[107] ——, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[108] B. Matérn, *Spatial variation*. Springer Science & Business Media, 2013, vol. 36.

[109] B. D. Ripley, *Spatial statistics*. John Wiley & Sons, 2005, vol. 575.

[110] N. Cressie, *Statistics for spatial data*. John Wiley & Sons, 2015.

[111] S. Meyer, L. Held, *et al.*, "Power-law models for infectious disease spread," *The Annals of Applied Statistics*, vol. 8, no. 3, pp. 1612–1639, 2014.

[112] R. Furrer, M. G. Genton, and D. Nychka, "Covariance tapering for interpolation of large spatial datasets," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 502–523, 2006.

[113] H. Rue and L. Held, *Gaussian Markov random fields: Theory and applications*. CRC press, 2005.

[114] H. Rue, S. Martino, and N. Chopin, "Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations," *Journal of the royal statistical society: Series b (statistical methodology)*, vol. 71, no. 2, pp. 319–392, 2009.

[115] H. RUE and H. Tjelmeland, "Fitting gaussian markov random fields to gaussian fields," *Scandinavian journal of Statistics*, vol. 29, no. 1, pp. 31–49, 2002.

[116] F. Lindgren, H. Rue, and J. Lindström, "An explicit link between gaussian fields and gaussian markov random fields: The stochastic partial differential equation approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 4, pp. 423–498, 2011.

[117] D. Simpson, F. Lindgren, and H. Rue, "Think continuous: Markovian gaussian models in spatial statistics," *Spatial Statistics*, vol. 1, pp. 16–29, 2012.

[118] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*, ACM, 2004, p. 29.

[119] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 259–302, 1986.

[120] Q. Liu and A. Ihler, "Distributed parameter estimation via pseudo-likelihood," *ArXiv preprint arXiv:1206.6420*, 2012.

[121] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017.

[122] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proceedings of the April 18-20, 1967, spring joint computer conference*, ACM, 1967, pp. 483–485.

[123] C. Beale, "Measuring rurality: Rural-urban continuum codes," *United States Department of Agriculture, Economic Research Service*, 2004.

[124] C. A. Gotway and L. J. Young, "Combining incompatible spatial data," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 632–648, 2002.

[125] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, "Cluster analysis of gene expression dynamics," *Proceedings of the National Academy of Sciences*, vol. 99, no. 14, pp. 9121–9126, 2002.